

BAB I

PENDAHULUAN

1.1 Latar Belakang

Tuberkulosis (TB) adalah penyakit menular yang menjadi masalah kesehatan dunia dan dapat berakibat fatal jika tidak segera diobati dengan benar. Penyakit yang disebabkan oleh infeksi bakteri *Mycobacterium Tuberculosis* ini umumnya menyerang paru-paru dan dapat menyebar melalui udara, menjadikannya salah satu penyakit paling mematikan di dunia [1]. Menurut laporan dari *World Health Organization (WHO)*, diperkirakan ada 1,25 juta kematian akibat TB pada tahun 2023 [2]. Angka ini lebih rendah dari tingkat pra-pandemi tahun 2019 yang mencapai 1,34 juta [2]. Indonesia sendiri merupakan salah satu negara dengan beban TB tertinggi, menempati peringkat kedua di dunia dengan estimasi 969.000 kasus pada tahun 2022 [3]. Tingginya angka prevalensi dan mortalitas ini menegaskan urgensi penanganan TB sebagai prioritas utama dalam Tujuan Pembangunan Berkelanjutan (SDGs) di bidang kesehatan [4].

Upaya pengendalian penyebaran TB sangat bergantung pada deteksi dini dan diagnosis yang akurat, yang memungkinkan pasien untuk segera mendapatkan pengobatan yang sesuai dan meminimalkan risiko komplikasi lebih lanjut [5]. Saat ini, fasilitas kesehatan tingkat pertama seperti puskesmas umumnya menggunakan metode pemeriksaan mikroskopis Bakteri Tahan Asam (BTA) pada spesimen dahak, namun metode manual ini memiliki kelemahan signifikan seperti memerlukan waktu pengerjaan yang panjang dan subjektivitas dalam penghitungan [4]. Di sisi lain, metode diagnostik konvensional berbasis kultur, yang merupakan standar baku, juga memiliki keterbatasan berupa waktu tunggu hasil yang lama (*extended turnaround times*), sehingga menghambat penanganan cepat [6]. Berbagai tantangan dalam penemuan kasus ini menyoroti pentingnya peran serta kader kesehatan di komunitas untuk membantu menjangkau populasi yang lebih luas [3].

Seiring dengan kemajuan teknologi, pemanfaatan *Computer Vision* dan *Machine Learning (ML)* menawarkan solusi inovatif untuk mengatasi berbagai tantangan dalam diagnosis medis [7]. *Machine learning* merupakan pengembangan

dari model statistika yang memungkinkan komputer untuk belajar dari data dan menghasilkan informasi baru secara otomatis [8]. Pendekatan ini telah terbukti andal dalam mendukung metode diagnosis, baik secara statis maupun dinamis [9]. Dalam konteks TB, teknologi *automated image analysis* telah dikembangkan untuk menghitung jumlah BTA dari citra mikroskopis [4], sementara metode *Deep Learning* seperti *Convolutional Neural Network (CNN)* juga telah banyak diterapkan untuk mendeteksi TB melalui citra medis seperti foto *rontgen* dada (*X-Ray*) dan *CT-Scan* dengan tingkat akurasi yang menjanjikan [10]. Beberapa penelitian menunjukkan bahwa model-model seperti *DenseNet121*, *ResNet50*, dan *Xception* mampu mencapai akurasi klasifikasi TB di atas 90% [1].

Selain data citra, data klinis dan rekam medis pasien merupakan sumber informasi yang kaya untuk diagnosis dan prediksi penyakit [6]. Algoritma *machine learning* dengan pendekatan *supervised learning* sangat efektif dalam menganalisis data terstruktur untuk tugas klasifikasi [9]. Berbagai algoritma populer seperti *Decision Tree*, *Random Forest*, *AdaBoost*, dan *Extreme Gradient Boosting (XGBoost)* telah digunakan secara luas untuk menganalisis berbagai kondisi kesehatan [5]. Algoritma-algoritma ini, terutama yang berbasis *ensemble learning*, memiliki kemampuan unggul dalam menangani data yang kompleks dan beragam, seperti yang ditunjukkan pada klasifikasi regimen obat anti-TB [11].

Di antara berbagai algoritma yang ada, *XGBoost* telah menunjukkan performa yang sangat baik dalam berbagai domain, termasuk klasifikasi medis. *XGBoost* adalah penyempurnaan dari algoritma *gradient boosting* yang dikenal efisien, memiliki ketahanan terhadap *outlier*, waktu komputasi yang singkat, dan mampu menghasilkan prediksi yang sangat akurat [12]. Sejumlah penelitian telah membuktikan keunggulan *XGBoost* dalam mendeteksi berbagai penyakit seperti gagal jantung [8], diabetes [13], dan kanker serviks [14]. Dalam beberapa studi perbandingan, *XGBoost* terbukti menghasilkan akurasi lebih tinggi dibandingkan algoritma lain seperti *SVM* dan *Random Forest*, bahkan mampu mencapai akurasi tertinggi hingga 97,21% setelah melalui proses seleksi fitur [15].

Meskipun memiliki performa dasar yang kuat, kinerja model *machine learning* sangat bergantung pada konfigurasi *hyperparameter* yang digunakan [16]. Penggunaan algoritma dengan pengaturan *default* seringkali tidak menghasilkan

performa yang optimal karena setiap dataset memiliki karakteristik yang unik [17]. Oleh karena itu, proses *hyperparameter tuning* menjadi langkah krusial untuk menemukan kombinasi *hyperparameter* terbaik yang dapat memaksimalkan kinerja model [18]. Berbagai penelitian telah menunjukkan bahwa *hyperparameter tuning* mampu memberikan peningkatan performa yang signifikan. Sebagai contoh, sebuah studi berhasil meningkatkan akurasi model *XGBoost* untuk prediksi gagal jantung dari 90,4% menjadi 94% [8], sementara studi lain pada klasifikasi anti-TB berhasil meningkatkan akurasi dari 90% menjadi 96% setelah optimasi menggunakan *Optuna* [11]. Peningkatan akurasi juga terlihat pada klasifikasi *COVID-19*, di mana akurasi validasi meningkat sebesar 17,1% setelah *hyperparameter tuning* diterapkan [10].

Selain tantangan dalam konfigurasi *hyperparameter*, penggunaan data medis klinis sering kali dihadapkan pada masalah ketidakseimbangan distribusi kelas (*class imbalance*) yang dapat mempengaruhi validitas diagnosis [19]. Dalam kondisi ini, algoritma klasifikasi cenderung menghasilkan bias dengan memprioritaskan prediksi pada kelas mayoritas demi mencapai akurasi semu, sementara gagal mengenali kelas minoritas yang justru merupakan target utama deteksi [20]. Analisis empiris menunjukkan bahwa kinerja algoritma *ensemble* seperti *XGBoost* dan *Random Forest* mengalami penurunan signifikan ketika dihadapkan pada berbagai tingkat ketidakseimbangan data tanpa penanganan yang tepat [21]. Oleh karena itu, penerapan metode *Synthetic Minority Over-sampling Technique* (SMOTE) yang diintegrasikan dengan *XGBoost* terbukti menjadi pendekatan yang efektif untuk meningkatkan ketahanan (*robustness*) model dan memperbaiki sensitivitas prediksi secara keseluruhan [22].

Berdasarkan tinjauan literatur, pemanfaatan *machine learning* untuk diagnosis TB telah banyak dilakukan, terutama menggunakan data citra dengan model *Deep Learning* [1]. Di sisi lain, algoritma *XGBoost* telah terbukti sangat efektif untuk klasifikasi pada berbagai data medis non-citra [23]. Namun, penelitian yang secara spesifik menerapkan algoritma *XGBoost* dengan proses *hyperparameter tuning* yang sistematis untuk klasifikasi penyakit tuberkulosis berdasarkan data klinis pasien masih terbatas. Penelitian sebelumnya seringkali berfokus pada perbandingan beberapa algoritma tanpa melakukan optimasi

mendalam atau menerapkan optimasi pada algoritma lain seperti *CatBoost* untuk klasifikasi regimen obat TB [11]. Adanya celah penelitian (*research gap*) ini memberikan peluang untuk mengembangkan sebuah model klasifikasi TB yang tidak hanya akurat, tetapi juga memiliki performa yang optimal.

Oleh karena itu, penelitian ini akan berfokus pada penerapan algoritma *XGBoost* dengan *hyperparameter tuning* untuk membangun model klasifikasi penyakit tuberkulosis. Penelitian ini bertujuan untuk mencari arsitektur model *XGBoost* terbaik dengan melakukan optimasi pada *hyperparameter* kunci untuk memaksimalkan akurasi prediksi. Hasil dari penelitian ini diharapkan dapat menghasilkan sebuah model klasifikasi dengan performa tinggi yang dapat diimplementasikan sebagai alat bantu dalam sistem pendukung keputusan klinis, membantu tenaga medis dalam melakukan diagnosis dini TB secara lebih cepat dan akurat, serta berkontribusi pada upaya penanggulangan tuberkulosis secara nasional dan global.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Metode diagnosis Tuberkulosis (TB) yang umum digunakan, seperti pemeriksaan BTA mikroskopis, masih memiliki kelemahan signifikan, yaitu prosesnya yang lama dan hasil yang bergantung pada penilai (subjektif). Hal ini menghambat upaya deteksi dini. Algoritma *machine learning* seperti *XGBoost* sebenarnya menawarkan solusi, namun kinerjanya sangat sensitif terhadap konfigurasi *hyperparameter* dan keseimbangan distribusi data (*data imbalance*). Menggunakan pengaturan default pada data klinis yang cenderung memiliki jumlah kasus positif lebih sedikit dibandingkan kasus negatif hampir pasti tidak akan memberikan hasil optimal dan berisiko menghasilkan prediksi yang bias ke kelas mayoritas. Masalah utamanya adalah masih jarangya penelitian yang mengaplikasikan *XGBoost* secara spesifik untuk klasifikasi TB dari data klinis, yang disertai dengan penanganan ketidakseimbangan data serta proses *hyperparameter tuning* yang sistematis. Tanpa langkah optimasi krusial ini, model yang dihasilkan tidak akan memuaskan, dan potensinya untuk menjadi alat bantu keputusan klinis yang cepat dan akurat dalam diagnosis TB menjadi tidak terwujud.

1.2.2 Solusi Permasalahan

Solusi untuk mengatasi performa model yang tidak optimal adalah dengan memperbaiki kualitas distribusi data latih serta tidak lagi bergantung pada parameter *default XGBoost*. Penelitian ini akan menerapkan metode *Synthetic Minority Over-sampling Technique (SMOTE)* untuk menangani ketidakseimbangan data, yang dilanjutkan dengan *hyperparameter tuning* yang sistematis menggunakan data klinis pasien dari RS Syarifah Ambami Rato Ebuh, Puskesmas, dan Kaggle. Metode optimasi yang digunakan adalah Grid Search, yang digabungkan dengan validasi silang (*K-Fold Cross-Validation*, $k = 10$) untuk mencari setelan parameter (*learning_rate*, *max_depth*, *n_estimators*, *gamma*, *min_child_weight*) yang paling efektif. Penelitian ini secara spesifik akan membandingkan kinerja model tanpa *tuning* dengan model hasil *tuning*, untuk menghasilkan sebuah model klasifikasi TB yang terbukti andal dan siap diimplementasikan sebagai alat bantu keputusan klinis.

1.2.3 Pertanyaan Penelitian

Berdasarkan latar belakang, rumusan masalah, dan solusi yang diusulkan, penelitian ini dirancang untuk menjawab pertanyaan-pertanyaan spesifik berikut:

1. Bagaimana tingkat performa model klasifikasi *XGBoost* dengan menggunakan parameter sebelum dilakukan *tuning* dalam mendeteksi penyakit tuberkulosis berdasarkan data klinis?
2. Seberapa signifikan peningkatan akurasi yang dicapai setelah dilakukan proses *hyperparameter tuning* dalam mendeteksi penyakit tuberkulosis berdasarkan data klinis?

1.3 Batasan Masalah

Untuk memastikan penelitian ini tetap fokus dan mendalam, perlu ditetapkan batasan-batasan yang jelas. Berikut adalah batasan masalah yang akan diterapkan dalam penelitian ini:

1. Penelitian ini secara spesifik berfokus pada penerapan dan optimasi algoritma *XGBoost*.
2. Data yang digunakan terbatas pada data klinis atau rekam medis pasien yang bersifat terstruktur (tabular) sebanyak 402 data yang bersumber dari RS Syarifah

Ambami Rato Ebuh periode 2023, 146 data dari Puskesmas, dan 10.000 data dari Kaggle. Penelitian ini tidak akan menggunakan atau melakukan pengolahan data citra medis, seperti foto *rontgen* dada, *CT-Scan*, ataupun citra mikroskopis BTA.

3. Proses optimasi model akan terkonsentrasi pada *hyperparameter tuning* untuk menemukan konfigurasi terbaik dari algoritma *XGBoost*, parameter yang digunakan adalah *n_estimators*, *max_depth*, *learning_rate*, *gamma*, *min_child_weight*.
4. Fitur yang digunakan dalam pemodelan terdapat 5 fitur input dan 1 target meliputi: Umur (Tahun), Jenis Kelamin (Laki-laki/Perempuan), Batuk Darah (Iya/Tidak), Sesak Nafas (Iya/Tidak), Sputum Test (Negative/Positive), dan Diagnosis (TBC/Bukan TBC).
5. Proses optimasi model akan menggunakan metode *Grid Search* untuk *hyperparameter tuning*, di mana pencarian konfigurasi terbaik dari algoritma *XGBoost* dilakukan secara sistematis dengan mencoba semua kemungkinan kombinasi *hyperparameter* yang telah ditentukan sebelumnya.
6. Model yang dikembangkan bertujuan untuk mengklasifikasikan status penyakit tuberkulosis dan tidak mencakup tugas prediksi lain seperti prognosis masa pengobatan, tingkat keparahan penyakit, atau prediksi resistansi obat.
7. Target atau output dari model klasifikasi dibagi menjadi 2 kelas, yaitu Tuberkulosis (TBC) dan Bukan Tuberkulosis (Bukan TBC).

1.4 Tujuan dan Manfaat

1.4.1 Tujuan

1. Menganalisis tingkat performa model klasifikasi *XGBoost* dengan menggunakan parameter sebelum dilakukan *tuning* dalam mendeteksi penyakit tuberkulosis berdasarkan data klinis.
2. Mengukur seberapa besar peningkatan akurasi yang dicapai model *XGBoost* setelah dilakukan proses *hyperparameter tuning* dalam mendeteksi penyakit tuberkulosis.

1.4.2 Manfaat

1. Memberikan referensi mengenai implementasi algoritma *XGBoost* yang dikombinasikan menggunakan *Grid Search* dan *SMOTE* untuk menangani data medis yang tidak seimbang.

2. Memberikan bukti tentang sejauh mana pengaruh *hyperparameter tuning* dalam meningkatkan performa model klasifikasi.
3. Hasil penelitian ini dapat dijadikan sebagai referensi tambahan bagi mahasiswa lain yang ingin mendalami penerapan teknik *ensemble learning* pada data medis.
4. Memudahkan pasien untuk melakukan deteksi dini mengenai gejala tuberkulosis dan memungkinkan pasien untuk segera mendapatkan pengobatan yang tepat, sehingga dapat meminimalkan risiko komplikasi lebih lanjut dan menekan potensi penularan penyakit.

1.5 Sistematika Penulisan

Adapun sistematika penulisan laporan proposal skripsi terdiri dari beberapa bagian, yaitu:

BAB I PENDAHULUAN

Membahas latar belakang, rumusan masalah, batasan, tujuan, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Menyajikan konsep, landasan teori, dan hasil penelitian sebelumnya yang relevan.

BAB III METODE PENELITIAN

Menjelaskan tahapan penelitian, arsitektur sistem, analisis kebutuhan, rancangan metode, dan skenario pengujian.

BAB IV HASIL PENELITIAN & ANALISA

Menjelaskan hasil penelitian dan analisa yang menjabarkan proses pengujian hingga menghasilkan output dari setiap tahapan pengujian.

BAB V PENUTUP

Memberikan kesimpulan sebagai jawaban dari rumusan masalah penelitian yang diperoleh serta memberikan saran sebagai referensi bagi penelitian selanjutnya.