

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pada tahun 2019, *Global Burden of Disease* mengeluarkan data mengenai prevalensi dan insiden stroke di seluruh dunia, hal ini telah disetujui oleh *World Stroke Organization (WSO)*. Secara *Global*, terdapat lebih dari 12,2 juta kasus *stroke* baru setiap tahunnya, dan satu dari empat orang di atas usia 25 tahun akan mengalami *stroke* dalam hidup mereka. Khususnya, lebih dari 16% dari semua kasus *stroke* terjadi pada usia 15-49 tahun dan lebih dari 62% terjadi pada mereka yang berusia di bawah 70 tahun. Dari segi distribusi jenis kelamin, 47% dari semua *stroke* terjadi pada pria dan 53% terjadi pada wanita [1]. Penyakit ini sangat berbahaya karena penyakit ini bisa berdampak ringan sampai sangat parah bagi penderita karena penderita bisa mengalami kerusakan sementara atau permanen. Faktor yang dapat meningkatkan untuk seseorang terkena *stroke* adalah penyakit serupa seperti memiliki infark miokard, pernah mengalami *stroke*, berbagai penyakit jantung seperti gagal jantung, fibrilasi atrium, usia lebih dari 55 tahun, hipertensi, stenosis karotis dari aterosklerosis, merokok, diabetes, kolesterol, darah tinggi, obesitas, gaya hidup menetap, konsumsi alkohol, terapi estrogen dan penggunaan zat euforia seperti kokain dan amfetamin [2].

Stroke dikenal sebagai penyebab utama kecacatan dan kematian di seluruh dunia, memberikan beban ekonomi yang substansial pada sistem kesehatan. Saat ini, sekitar 34% dari total kesehatan global dialokasikan untuk perawatan *stroke*, dengan biaya rata-rata per pasien *stroke* di Amerika Serikat mencapai perkiraan USD 140.048 yang meliputi biaya rawat inap, rehabilitasi, dan tindak berkelanjutan. Secara mengkhawatirkan, telah terjadi peningkatan signifikan

dalam insiden stroke di kalangan populasi yang lebih muda, terutama pada negara-negara berpendapatan rendah hingga menengah, insiden pada individu berusia 20 sampai 64 tahun telah meningkat sebesar 25% [3]. Di luar biaya medis langsung, *stroke* juga menambah beban biaya produktivitas nasional karena pasien sering mengalami disabilitas membatasi kemandirian yang mempengaruhi produktivitas keluarga yang lain. [4]. Biaya sosial yang harus ditanggung oleh penderita *stroke* cukup tinggi dihitung dari biaya rawat inap untuk perawatan akut dan rehabilitasi awal, serta janji klinik berikutnya, resep, perawatan jangka panjang, biaya terapi, dan hilangnya pendapatan karena ketidakmampuan penderita *stroke* untuk bekerja [5].

Mengingat banyaknya korban dari penyakit *stroke* ini membuat para peneliti telah mengembangkan berbagai teknik otomatis untuk mendeteksi *stroke* dan mengetahui tingkat risiko dari penyakit *stroke*. Hal ini menjadi perhatian peneliti untuk mengetahui cara mengantisipasi penyakit *stroke* karena tingkat risikonya tinggi. Solusi yang dapat diberikan dapat diketahui lewat analisa terhadap risiko penyakit *stroke* melalui pendekatan *data mining*. Pada penelitian sebelumnya yang digunakan sebagai acuan dalam penelitian ini. Penelitian pertama [6] mengevaluasi efektivitas metode *10-Cross Fold Validation* dan metode pemisahan data 80:20 dalam memprediksi hujan di Australia menggunakan *Decision Tree*. Akurasi menggunakan *10-Cross Fold Validation* pada algoritma *Decision Tree* memperoleh hasil tertinggi sebesar 87,35% dibandingkan dengan 85,37% dari metode pemisahan data dalam memprediksi hujan di Australia. Penelitian kedua [7] menguji dua model prediksi penyakit kardiovaskular dengan dan tanpa menggunakan *Particle Swarm Optimization (PSO)* sebagai algoritma pemilihan fitur pada *Decision Tree*. Model dengan pemilihan fitur memperoleh akurasi sebesar 85,24% dengan waktu pemrosesan 0,11 detik, sedangkan model tanpa pemilihan fitur memperoleh akurasi sebesar 83,61% dengan waktu pemrosesan 0,02 detik. Penelitian ketiga [8]

menggunakan algoritma *PSO* sebagai seleksi fitur pada tiga algoritma klasifikasi yaitu *Xgboost*, *Naïve Bayes*, dan *ID3* untuk mendeteksi tumor otak berdasarkan gambar. Algoritma yang menunjukkan performa terbaik adalah *PSO* pada *Xgboost* dengan akurasi 97%, presisi 98%, dan sensitivitas 97%. Penelitian keempat [9] menggunakan *Ant Colony Optimization (ACO)* sebagai seleksi fitur pada algoritma *Xgboost* untuk identifikasi emitor tertentu. *ACO-Xgboost* mampu meningkatkan akurasi pada tiga *dataset* berbeda, yaitu *dataset* asli, *dataset* 10db, dan *dataset* 5db, sebesar 0,20%-3,53% dibandingkan dengan algoritma lain yang digunakan untuk klasifikasi. Penelitian kelima [10] menggunakan pengklasifikasi *Extra Tree Classifier* guna mengidentifikasi lokasi dan jenis patah tulang dalam gambar. Akurasi pengklasifikasi *Extra Tree* menunjukkan kinerja yang baik dengan akurasi 92,13%, presisi 91,85%, dan sensitivitas 90,96%. Belum ada penelitian lebih lanjut yang melakukan perbandingan tiga algoritma berbasis pohon keputusan dengan seleksi fitur menggunakan *PSO* dan *ACO*.

Pemilihan algoritma berbasis pohon keputusan untuk klasifikasi didasari oleh performa algoritma dengan membandingkan algoritma klasifikasi *machine learning* lain seperti *Linear SVM* yang memiliki akurasi 0.7633 dengan waktu pemrosesan 3.42 detik, *K-Nearest Neighbors* memiliki akurasi 0.8622, presisi 0.8913, dan sensitivitas 0.8627 dengan waktu pemrosesan 116.9920 detik. Akurasi dan presisi *KNN* cukup baik namun waktu pemrosesannya yang lama membuatnya kurang optimal. *Naive Bayes* memiliki akurasi 0.7302 dengan waktu pemrosesan 0.4242 detik. Meskipun waktu pemrosesannya sangat cepat namun akurasinya rendah, *Logistic Regression* memiliki akurasi 0.7632 dengan waktu pemrosesan 1.2850 detik. Seperti *Linear SVM*, *Logistic Regression* cepat dan sederhana, tetapi hasilnya tidak optimal. Hasilnya algoritma *machine learning* lain tidak sebaik algoritma berbasis pohon keputusan yang menunjukkan potensi hasil akurasi yang berada di atas 85% dan berdasarkan

hasil penelitian sebelumnya [6], [11], [12], [7], [13], [14], [15], [8], [9], [10] dengan algoritma *machine learning* untuk melakukan klasifikasi menggunakan algoritma seperti *Decision Tree*, *Extra tree classifier* dan *Xgboost* termasuk dengan melakukan metode tambahan seperti seleksi fitur menggunakan *ACO* dan *PSO*, belum terdapat penelitian lebih lanjut untuk pembuatan komparasi dari ketiga algoritma tersebut yang berbasis pohon keputusan dengan seleksi fitur menggunakan *PSO* dan *ACO*. Penelitian sebelumnya telah menunjukkan peningkatan efektivitas dengan menggunakan seleksi fitur dan hasil akurasi yang tinggi dari algoritma konvensional itu sendiri.

Kontribusi pada penelitian ini adalah 1) mengetahui penerapan algoritma berbasis pohon keputusan untuk klasifikasi penyakit *stroke* 2) data yang digunakan pada penelitian ini adalah *indicator of heart disease (2022)*, data yang menyediakan variabel penyebab berbagai penyakit kardiovaskular 3) menerapkan model klasifikasi untuk penyakit *stroke* 4) mengoptimalkan model algoritma berbasis pohon keputusan dengan seleksi fitur menggunakan *ACO* dan *PSO* 5) melakukan perbandingan antara ketiga algoritma berbasis keputusan dengan dan tanpa seleksi fitur menggunakan *PSO* dan *ACO* untuk menemukan model dengan hasil terbaik untuk klasifikasi penyakit *stroke*.

1.2 Rumusan Masalah

Rumusan Masalah dari penelitian ini di antaranya:

- 1) Bagaimana menerapkan algoritma berbasis pohon keputusan yaitu *Decision Tree*, *Extra tree classifier*, dan *Xgboost* dalam klasifikasi penyakit *stroke* dilihat dari hasil akurasi, sensitivitas, presisi, dan *time processing*?
- 2) Bagaimana menerapkan algoritma optimasi yang berbasis *swarm intelligence* yaitu *PSO* dan *ACO* sebagai seleksi fitur pada klasifikasi penyakit *stroke* berbasis pohon keputusan?

- 3) Bagaimana hasil akurasi, sensitivitas, presisi, dan *time processing* dari algoritma *PSO* dan *ACO* untuk mengoptimalkan algoritma berbasis pohon keputusan dalam klasifikasi penyakit *stroke*?

1.3 Batasan Masalah

Batasan masalah dari penelitian ini di antaranya:

- 1) *Dataset Indicators of Heart Disease* tahun 2022 diambil dari *website kaggle* dengan 40 variabel dan 445133 observasi.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

Tujuan dari penelitian ini di antaranya:

- 1) Menerapkan algoritma berbasis *pohon keputusan* yaitu *Decision Tree*, *Extra tree classifier*, dan *Xgboost* untuk mengklasifikasikan penyakit *stroke*.
- 2) Menerapkan seleksi fitur dengan algoritma *Swarm Intelligence* untuk mengoptimalkan algoritma berbasis *pohon keputusan* dalam mengklasifikasikan penyakit *stroke*.
- 3) Melakukan analisa perbandingan terhadap hasil akurasi, sensitivitas, presisi, dan *time processing* dari algoritma berbasis *pohon keputusan* yang sudah menggunakan fitur terpilih hasil dari proses seleksi fitur dengan algoritma *Swarm Intelligence* yaitu *PSO* dan *ACO*.

1.4.2 Manfaat Penelitian

Manfaat teoritis dari penelitian ini di antaranya:

- 1) Manfaat teoritis pada penelitian ini adalah memberikan kontribusi kepada peneliti selanjutnya mengenai efektivitas teknik *machine learning* dengan algoritma optimasi.
- 2) Menganalisis fitur yang dipilih oleh algoritma optimasi, peneliti dan praktisi dapat memperoleh wawasan tentang fitur-fitur yang paling penting dalam klasifikasi penyakit *stroke*.

Manfaat praktis pada penelitian ini adalah untuk memberikan wawasan bagi pembaca untuk meningkatkan kewaspadaan terhadap penyakit *stroke* berdasarkan faktor yang mempengaruhi yang dapat mengakibatkan cacat ringan sampai permanen kepada penderitanya.

1.5 Sistematika Penulisan

BAB 1 PENDAHULUAN

Berisi pengantar dan pengenalan mengenai bahaya dari penyakit *stroke* sebagai wawasan dasar dari penelitian ini.

BAB 2 LANDASAN TEORI

Menjelaskan terkait studi literatur terkait seluruh teori, *framework* yang digunakan, jurnal, *tools*, dan algoritma yang akan digunakan dan menjadi landasan penelitian ini.

BAB 3 METODELOGI PENELITIAN

Membahas mengenai gambaran besar objek penelitian, metodologi, seluruh langkah-langkah yang dilakukan pada penelitian ini dengan tujuan agar penelitian ini memberikan solusi dari permasalahan yang diambil.

BAB 4 ANALISIS DAN HASIL PENELITIAN

Membahas mengenai tahapan yang dilakukan menggunakan metode *data mining* yang ditentukan mulai dari *Business understanding* sampai *Evaluation* dan hasil dan diskusi.

BAB 5 SIMPULAN

Membahas Simpulan dengan cara menjawab rumusan masalah yang telah ditentukan sebelumnya. Memberikan saran kepada peneliti selanjutnya yang tertarik dengan klasifikasi *stroke*.