BAB II

LANDASAN TEORI

2.1 Tinjauan Pustaka

2.1.1 Autisme dan Karakteristiknya

Autisme atau Autism Spectrum Disorder (ASD) adalah gangguan perkembangan saraf yang memengaruhi kemampuan individu dalam berkomunikasi, berinteraksi sosial, serta menunjukkan pola perilaku dan minat yang terbatas dan berulang. Gangguan ini biasanya muncul sejak usia dini dan berlangsung seumur hidup. (Sukma, 2023) menjelaskan bahwa anak dengan autisme memiliki tantangan dalam menanggapi lingkungan sosial, misalnya tidak menunjukkan kontak mata, tidak merespon ketika dipanggil, atau mengalami kesulitan memahami bahasa tubuh. Ciri lainnya adalah kecenderungan melakukan aktivitas tertentu secara berulang dan ketertarikan yang terbatas pada objek atau rutinitas tertentu.

Karakteristik perkembangan anak dengan autisme juga dipengaruhi oleh tingkat keparahan gangguan. menunjukkan bahwa anak dengan autisme ringan hingga sedang masih dapat mengikuti pembelajaran secara sistematis, sedangkan anak dengan autisme berat cenderung membutuhkan bantuan penuh dari guru untuk fokus dan memahami instruksi. Sementara itu, (Shinta Delfianti et al., 2024) menyoroti bahwa anak autis memiliki tantangan dalam komunikasi ekspresif dan reseptif, serta menunjukkan pola interaksi yang minim dengan lingkungan sosialnya. Pemahaman terhadap karakteristik ini sangat penting dalam penyusunan

sistem intervensi atau alat bantu yang tepat, termasuk penerapan teknologi untuk mendeteksi atau memfasilitasi kebutuhan mereka.

2.1.2 Data Mining

Data mining adalah proses yang digunakan untuk mengekstraksi informasi yang berguna dari kumpulan data yang besar. Proses ini mencakup teknik-teknik statistik, algoritmik, dan pembelajaran mesin guna menemukan pola tersembunyi, anomali, serta hubungan dalam data. Menurut (Blum et al., 2022), data mining sangat penting dalam dunia modern karena dapat digunakan untuk mengambil keputusan yang berbasis data secara efisien, khususnya ketika data bersifat kompleks dan berdimensi tinggi. Dalam konteks pendidikan, data mining telah digunakan untuk menganalisis performa siswa, mendeteksi risiko putus sekolah, hingga mengevaluasi efektivitas sistem pembelajaran.

Dalam penelitian ini, data mining diterapkan untuk menganalisis data kuisioner gejala autisme siswa SLB. Melalui tahapan standar seperti pembersihan data dan pemodelan, pendekatan ini membantu membangun sistem prediksi berbasis web yang objektif untuk deteksi dini autisme.

2.1.3 Machine Learning

Machine learning adalah cabang dari kecerdasan buatan (AI) yang berfokus pada pengembangan algoritma dan teknik yang memungkinkan komputer untuk “belajar” dari data. Algoritma machine learning bekerja dengan mengenali pola dalam data dan kemudian menggunakannya untuk membuat prediksi atau keputusan secara otomatis tanpa diprogram secara eksplisit untuk tugas tersebut.

Menurut (Géron, 2022), machine learning sangat penting dalam sistem yang membutuhkan prediksi berbasis data, termasuk dalam bidang pendidikan dan kesehatan, karena mampu menangani kompleksitas variabel yang tinggi dengan efisien.

Terdapat berbagai pendekatan dalam machine learning, salah satunya adalah supervised learning, di mana algoritma dilatih dengan data berlabel. Salah satu algoritma yang populer adalah Random Forest, yang terdiri dari kumpulan decision tree dan mampu memberikan hasil klasifikasi yang stabil dan akurat. Dalam penerapan nyata, seperti prediksi indikasi autisme pada siswa SLB, algoritma ini sangat berguna karena mampu mempertimbangkan banyak fitur sekaligus dan mengurangi risiko overfitting yang umum pada pohon keputusan Tunggal.

2.1.4 Algoritma Random Forest

Random Forest adalah salah satu metode ensemble learning yang membangun banyak pohon keputusan (decision trees) selama proses pelatihan, lalu menggabungkan hasilnya (baik untuk klasifikasi maupun regresi) melalui voting atau rata-rata. Seperti dijelaskan oleh (Géron, 2022), Random Forest memanfaatkan prinsip bagging (bootstrap aggregating), di mana setiap pohon dilatih pada subset acak dari data dan fitur, sehingga menghasilkan model yang kuat terhadap overfitting serta lebih stabil dibanding satu pohon tunggal

(Alex J. Smola & S.V.N. Vishwanathan, 2022) menekankan bahwa pendekatan ensemble seperti Random Forest bekerja baik dalam situasi di mana hubungan antara fitur tidak linier atau sulit dimodelkan secara eksplisit. Dengan

membentuk banyak model lemah (weak learners) yang berbeda dan mengombinasikannya, Random Forest mampu menangkap pola yang lebih kaya dalam data dibandingkan model sederhana. Selain itu, metode ini memberikan fitur penting dalam bentuk feature importance, yang membantu dalam interpretasi hasil dan pemilihan atribut yang relevan

2.1.5 CRISP-DM sebagai Tahapan Data Mining

CRISP-DM adalah model proses standar yang digunakan dalam proyek data mining. Model ini mencakup enam fase utama yang terstruktur secara sistematis dan iteratif. Pertama adalah Business Understanding, yaitu memahami tujuan bisnis dan bagaimana data mining bisa membantu mencapainya. Kedua, Data Understanding, yaitu mengumpulkan data awal dan mengenal karakteristiknya. Ketiga adalah Data Preparation, yang mencakup pembersihan, seleksi, dan transformasi data sebelum dimasukkan ke dalam model.

Tahap keempat adalah Modeling, di mana algoritma seperti Random Forest diterapkan untuk membangun model prediktif. Tahap kelima adalah Evaluation, yaitu menilai apakah model yang dibangun sudah memenuhi tujuan bisnis. Terakhir adalah Deployment, di mana model digunakan secara aktual, misalnya melalui aplikasi web berbasis sistem pendukung keputusan. Struktur ini membuat CRISP-DM sangat cocok untuk penelitian berbasis prediksi, seperti mendeteksi indikasi autisme menggunakan data screening siswa.

2.1.6 Explainable AI

Explainable AI (XAI) adalah bidang dalam kecerdasan buatan yang terdiri dari serangkaian metode dan teknik untuk memastikan keputusan dan hasil dari model *machine learning* dapat dipahami oleh manusia (Sopiandi et al., 2025). Tujuan utamanya adalah mengubah model yang bersifat "kotak hitam" (*black box*), di mana logikanya tidak transparan, menjadi model yang prosesnya dapat diinterpretasikan (Puspanagara, 2025). Pentingnya XAI terletak pada kemampuannya untuk meningkatkan kepercayaan (*trust*), memvalidasi cara kerja sistem, mempermudah proses *debugging* saat terjadi kesalahan, serta mendeteksi dan memitigasi potensi bias dalam model untuk menjamin keadilan (*fairness*) (Molnar, 2022).

2.1.7 SHAP

Dalam perkembangannya, terdapat berbagai metode untuk mencapai tujuan XAI, di mana SHAP dan LIME adalah yang paling populer (Sopiandi et al., 2025). SHAP (SHapley Additive exPlanations) adalah sebuah metode post-hoc yang didasarkan pada konsep Nilai Shapley (Shapley Value) dari teori permainan kooperatif (cooperative game theory) (Molnar, 2022).

Dalam teori permainan, Nilai Shapley adalah metode untuk mendistribusikan "pembayaran" secara adil kepada setiap "pemain" yang telah bekerja sama dalam sebuah permainan. Dalam konteks machine learning, "pemain" adalah fitur-fitur dari data, "permainan" adalah proses prediksi untuk satu instance, dan "pembayaran" adalah prediksi itu sendiri. SHAP menggunakan konsep ini

untuk menghitung seberapa besar kontribusi setiap fitur dalam "mendorong" prediksi dari nilai dasar (rata-rata prediksi dari seluruh dataset) ke prediksi akhir untuk suatu instance data (Molnar, 2022).

Hasilnya adalah nilai SHAP untuk setiap fitur, di mana nilai positif menandakan fitur tersebut meningkatkan prediksi, dan nilai negatif menurunkannya. Keunggulan utama SHAP adalah kemampuannya memberikan penjelasan lokal (untuk satu prediksi spesifik) dan penjelasan global (pentingnya fitur secara keseluruhan) secara konsisten dan akurat secara matematis (Molnar, 2022).

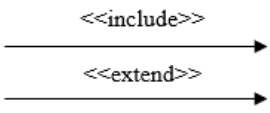
2.1.8 Pemodelan Sistem UML

Menurut (Pratama et al., 2023) Bahasa modeling visual Unified Modeling Language (UML) digunakan untuk mendokumentasikan, merancang, dan membangun sistem perangkat lunak. UML membantu tim pengembang perangkat lunak dalam merancang dan mengembangkan sistem yang kompleks dengan menggambarkan struktur dan perilaku sistem secara visual. Notasi UML yang standar dan mudah dipahami membuatnya mudah digunakan oleh pengembang perangkat lunak, manajer proyek, pengguna akhir, dan siapa pun yang ingin membuat sistem perangkat lunak.Beberapa jenis diagram UML yang umum digunakan mewakili berbagai aspek sistem perangkat lunak, seperti struktur sistem, interaksi antar objek, perilaku, dan proses bisnis, Beberapa jenis diagram UML yang umum digunakan adalah:

2.1.8.1 Use Case Diagram

Use case diagram adalah deskripsi suatu fungsi yang ada dalam sebuah sistem jika dilihat dari pandangan pengguna sistem. Sedangkan cara unjuk kerja use case diagram adalah menjelaskan dasar dari interaksi diantara pengembang dalam sebuah sistem dengan sistem itu sendiri. Use case diagram menjelaskan apa saja yang sistem akan lakukan, menjelaskan apa yang akan diperbuat oleh sistem dan tidak bagaimana. Menjelaskan fungsionalitas yang diharapkan dari sistem. Selain itu juga menggambarkan sebuah sistem dari pandangan pengguna

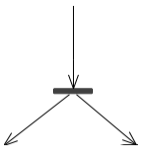
Tabel 2. 1 Komponen Use Case Diagram

Simbol	Nama Komponen	Fungsi
	Aktor (*Actor*)	untuk memberikan input ke sistem, menerima informasi dari sistem, atau melakukan keduanya—mengirim dan menerima informasi.
	Dependency	Untuk menghubungkan antara dua elemen, di mana perubahan pada satu elemen memengaruhi elemen lainnya
	Association	hubungan antara elemen-elemen struktural yang saling terhubung dalam sebuah sistem.
	Generalization	Untuk menggambarkan relasi khusus antara objek anak (*child*) dan induk (*parent*).
	Include dan extends	include menggambarkan perilaku yang wajib dipenuhi agar suatu *event* dapat terjadi extends menunjukkan perilaku tambahan yang hanya terjadi di bawah kondisi tertentu

2.1.8.2 Activity Diagram

Diagram aktivitas menunjukkan bagaimana pernyataan dilakukan dan disajikan dalam prosedur atau kinerja suatu aktivitas dalam alir kerja. Status tindakan, yang mirip dengan aktivitas, dapat dimasukkan ke dalam diagram aktivitas. Aktivitas diagram harus sejajar horizontal dengan jenis pemodelan lainnya, seperti diagram use case. Agar dapat membuat pemodelan dengan model alur kerja sistem yang baik, pengguna dapat menggunakan activity diagram.

Tabel 2. 2 Komponen Activity Diagram

Simbol	Nama Komponen	Fungsi
	• *Start State* • *End State*	• Sebagai tanda awal proses dari *activity diagram*. • menggambarkan akhir atau terminal dari pada sebuah activity diagra
	*Activity*	mengambarkan sebuah pekerjaan atau tugas dalam workflow
	*Decision*	mengindikasikan suatu kondisi kemungkinan perbedaan transisi
	State Transition	menunjukan kegiatan apa berikutnya setelah suatu kegiatan sebelumnya
	*Fork/percabangan*	menunjukkan kegiatan yang dilakukan secara paralel atau untuk menggabung kan dua kegiatan paralel menjadi satu

	*Join (penggabungan)*	menunjukkan adanya dekomposisi

2.1.9 Database Dan DBMS

Database Management System sendiri, yaitu sebuah software atau sistem yang dibuat dengan tujuan untuk membantu mengolah basis data dan mengeksekusi suatu operasi apabila diminta oleh banyak pengguna basis data (database user). Memastikan supaya basis data selalu terorganisir secara stabil dan bisa diakses dengan mudah adalah tugas dari DBMS yang merupakan sebagai penghubung antara basis data dengan application program. Jadi, dapat dikatakan bahwa DBMS akan menangani semua akses ke basis data yang dilakukan oleh user. Selain itu, DBMS juga berfungsi sebagai alat untuk mendefinsikan data, melakukan check terhadap keamanan dan integritas data yang diterjemahkan oleh DBA (Database Administrator), mengatasi hal-hal yang dianggap gagal dalam melakukan akses data, dan lain-lain.

2.1.10 Perangkat Lunak yang Digunakan

Penelitian modern di bidang *machine learning* dan *Explainable AI* sangat bergantung pada ekosistem perangkat lunak dan *library* yang matang. Bagian ini menjelaskan beberapa perangkat lunak dan teknologi utama yang menjadi fondasi dalam pengembangan sistem cerdas yang dapat diinterpretasikan.

2.1.10.1 Visual Studi Code

Visual Studio Code (VS Code) merupakan editor kode sumber terbuka yang dikembangkan oleh Microsoft, dan menjadi salah satu editor yang paling populer di kalangan pengembang perangkat lunak. VS Code mendukung berbagai bahasa pemrograman seperti Python, JavaScript, C++, HTML, dan banyak lagi. Kelebihan utama dari VS Code terletak pada antarmuka yang ringan, kustomisasi yang fleksibel, serta integrasi yang baik dengan berbagai ekstensi dan alat bantu pengembangan.

Dalam pengembangan aplikasi berbasis machine learning, termasuk sistem prediksi indikasi autisme, VS Code digunakan sebagai lingkungan pengembangan utama. Dengan dukungan ekstensi seperti Python, Jupyter, dan Live Server, pengguna dapat menulis kode, melakukan debugging, menjalankan skrip, hingga mengelola proyek secara efisien dalam satu platform. Selain itu, integrasi dengan Git memudahkan pengelolaan versi kode. VS Code juga memiliki fitur IntelliSense yang mendukung auto-completion dan dokumentasi instan, sangat membantu dalam proses pengembangan berbasis pustaka seperti scikit-learn atau TensorFlow.

2.1.10.2 MYSQL

MySQL adalah RDBMS sumber terbuka dan memiliki kemampuan untuk mengerjakan pekerjaan secara sekaligus. Pencipta MySQL, yaitu Michael “Monty” Widenius tahun 1995. Di tahun 2000, MySQL diumumkan dengan lisensi ganda, di mana mengizinkan semua orang menggunakannya secara free di bawah Lisensi Publik Umum GNU, sehingga kepopulerannya semakin memuncak. MySQL AB (AB = aktiebolag) merupakan perusahaan pemilik dan pengembang MySQL, di

mana perusahaan tersebut saat ini menjadi anak perusahaan dari Sun Microsystems. MySQL bisa menjadi basis data sepopuler seperti sekarang adalah bukan karena tanpa alasan, melainkan karena mempunyai kehebatan terhadap fitur fiturnya yang sangat beragam, salah satu fitur yang paling terkenal adalah masalah speed atau kecepatannya. E-Week melakukan perbandingan terhadap sejumlah database seperti MySQL, MS SQL, Oracle, Sybase ASE, dan IBM DB2. Dibuktikan, yaitu MySQL dan Oracle memiliki kemampuan terbaik dalam hal performance dan scalability. MySQL dapat dengan lancar dan cepat mengatasi tabel yang berjumlah puluhan ribu serta record data yang berjumlah miliaran.

2.1.10.3 Bahasa Pemrograman Python dan Library Terkait

Python telah menjadi bahasa pemrograman dominan dalam bidang machine learning dan analisis data karena beberapa keunggulan mendasar. Sebagai bahasa tingkat tinggi dengan sintaks yang intuitif dan mudah dipelajari, Python memungkinkan peneliti untuk fokus pada penyelesaian masalah daripada menghabiskan waktu untuk masalah teknis pemrograman. (Virtanen et al., 2020) menjelaskan bahwa Python, didukung oleh ekosistem perpustakaan yang kaya, telah merevolusi komputasi ilmiah dengan membuat analisis data kompleks menjadi lebih mudah diakses dan efisien. Fleksibilitas Python dalam menangani berbagai tugas, mulai dari pemrosesan data hingga pembangunan model prediktif, menjadikannya pilihan utama dalam penelitian ini.

Untuk kebutuhan manipulasi data tabular, penelitian ini memanfaatkan library Pandas. (McKinney, 2022) dalam bukunya menjelaskan bahwa Pandas

menyediakan struktur data DataFrame yang memungkinkan operasi seperti pembersihan data, transformasi, dan agregasi dilakukan dengan sintaks yang ringkas namun ekspresif. Kemampuan Pandas dalam menangani data missing value, melakukan merge dan join dataset, serta operasi time series sangat mendukung tahap pra-pemrosesan data dalam penelitian ini.

Sementara untuk komputasi numerik, NumPy menjadi pilihan utama. (Harris et al., 2020) menyoroti bahwa NumPy tidak hanya menyediakan array multidimensi yang efisien, tetapi juga menjadi fondasi bagi banyak library ilmiah Python lainnya. Optimasi yang dilakukan NumPy pada operasi vektor dan matriks memungkinkan komputasi numerik berjalan dengan performa tinggi, yang sangat krusial dalam proses pelatihan model machine learning.

Visualisasi data sebagai bagian penting dalam analisis eksploratori dan presentasi hasil dilakukan menggunakan Matplotlib dan Seaborn. (Waskom, 2021) mengembangkan Seaborn di atas Matplotlib untuk menyediakan antarmuka tingkat tinggi yang lebih intuitif dalam membuat visualisasi statistik yang menarik secara visual. Kombinasi kedua library ini digunakan dalam penelitian untuk mengeksplorasi distribusi data, hubungan antar variabel, serta mempresentasikan hasil prediksi model.

Scikit-learn adalah salah satu library machine learning paling populer di Python. Library ini menyediakan berbagai algoritma klasifikasi, regresi, klastering, serta alat evaluasi model. library machine learning Python yang menyediakan alat untuk pra-pemrosesan data, pelatihan model, dan evaluasi.

(Buitinck et al., 2023) menekankan bahwa scikit-learn tetap menjadi pilihan utama untuk implementasi model supervised learning seperti Decision Tree Regression karena stabilitas dan optimasinya. Dalam penelitian ini, modul sklearn.tree. DecisionTreeRegressor digunakan untuk membangun model prediktif.

2.1.10.4 Bahasa Pemrograman PHP

PHP adalah sebuah bahasa pemrograman server-side berbasis kode-kode (script) yang digunakan untuk mengolah data dan kemudian diproses di server. Jenis server yang paling umum digunakan antara lain Apache dan Nginx. Karena PHP bersifat open source, pengguna dapat mengembangkan dan mengubahnya sesuai keinginan mereka. PHP biasanya digunakan untuk membuat aplikasi komputer dan membuat website (dinamis atau statis). PHP mendukung berbagai protokol penting seperti POP3,IMAP, dan LDAP. Ini dapat digunakan bersama dengan berbagai database populer seperti MySQL, Oracle, Sybase, dan Microsoft SQL Server.

2.2 Kajian Penelitian Terdahulu

2.2.1 Paper 1

Penelitian yang dilakukan oleh Novianto dan Anasanti (2023) membahas identifikasi gangguan spektrum autisme (ASD) menggunakan pendekatan machine learning berbasis fitur. Penelitian ini memanfaatkan kombinasi tiga dataset dari UCI dengan total 1.100 individu dan menerapkan berbagai algoritma klasifikasi seperti

K-Nearest Neighbor (KNN), Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), serta Decision Tree (DT). Teknik imputasi seperti linear regression, MiceForest, dan MissForest digunakan untuk menangani data yang hilang, sedangkan teknik seleksi fitur seperti SpFSR, Mutual Information, dan Random Forest Importance diterapkan untuk memilih fitur paling relevan. Hasil penelitian menunjukkan bahwa kombinasi imputasi linear, pemilihan fitur SpFSR, dan algoritma SVM mampu memberikan akurasi prediksi tertinggi, yaitu mencapai 100%, yang mengungguli penelitian sebelumnya dalam bidang deteksi ASD

2.2.1.1 Tujuan Penelitian

Untuk mengembangkan model klasifikasi berbasis machine learning, termasuk Random Forest, dalam mengidentifikasi individu dengan dan tanpa ASD (Autism Spectrum Disorder), serta menguji efektivitas teknik imputasi dan seleksi fitur dalam meningkatkan akurasi prediksi.

2.2.1.2 Metodologi yang Digunakan

Data diambil dari gabungan tiga dataset UCI Autism Screening (N = 1.100). Penelitian menggunakan beberapa metode imputasi (Linear Regression, MissForest, MICE) dan seleksi fitur (spFSR, Mutual Information, F-Score, RFI). Metode klasifikasi yang diterapkan termasuk Random Forest, Support Vector Machine, KNN, Naive Bayes, dan Decision Tree. Evaluasi dilakukan dengan validasi silang 10-fold.

2.2.1.3 Temuan Utama

Random Forest menunjukkan performa sangat baik dengan akurasi mencapai 100% pada data hasil imputasi. Meskipun SVM menghasilkan akurasi tertinggi secara konsisten, Random Forest juga mampu mengklasifikasikan ASD secara akurat baik menggunakan semua fitur maupun subset fitur yang dipilih.

2.2.1.4 Kesimpulan Penelitian

Random Forest terbukti efektif sebagai salah satu algoritma klasifikasi ASD dengan tingkat akurasi tinggi. Model ini juga robust terhadap variasi data, terutama jika dikombinasikan dengan teknik imputasi dan seleksi fitur seperti spFSR. Hal ini menunjukkan Random Forest cocok digunakan untuk sistem prediksi ASD yang efisien dan handal.

2.2.2 Paper 2

Penelitian oleh Yuliani (2022) menerapkan algoritma Random Forest untuk memprediksi kelangsungan hidup pasien gagal jantung. Penelitian ini menggunakan dataset publik dari Kaggle dengan total 299 data pasien dan menerapkan metode seleksi fitur BestFirst yang menghasilkan empat fitur utama: usia, enjection fraction, serum creatinine, dan waktu pengamatan. Penanganan ketidakseimbangan kelas dilakukan dengan model class balancer. Dengan metode percentage split 80%, model Random Forest menghasilkan akurasi sebesar 91,45% dan nilai AUC sebesar 0,953. Hasil ini menunjukkan bahwa Random Forest efektif dalam menangani prediksi pada data medis, serta seleksi fitur yang tepat mampu meningkatkan performa model prediksi secara signifikan

2.2.2.1 Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk memprediksi kelangsungan hidup pasien penderita gagal jantung dengan menerapkan algoritma machine learning, khususnya algoritma Random Forest, serta melakukan seleksi fitur untuk meningkatkan akurasi prediksi.

2.2.2.2 Metodologi Yang Digunakan

Penelitian ini menggunakan tahapan CRISP-DM, yaitu: pemahaman bisnis, pemahaman data, persiapan data, pemodelan, dan evaluasi. Algoritma yang digunakan adalah Random Forest, Random Subspace, dan LogitBoost. Untuk meningkatkan performa, dilakukan seleksi fitur menggunakan metode BestFirst dan penyeimbangan data dengan model Class Balancer. Evaluasi dilakukan menggunakan metode cross validation dan percentage split (80%).

2.2.2.3 Temuan Utama

Penelitian ini menemukan bahwa fitur-fitur paling berpengaruh terhadap prediksi kelangsungan hidup adalah: age, ejection_fraction, serum_creatinine, dan time. Hasil terbaik diperoleh dengan algoritma Random Forest menggunakan percentage split 80%, dengan akurasi 91,45%, precision 0,915, recall 0,914, dan AUC 0,953.

2.2.2.4 Kesimpulan Penelitian

Penelitian menunjukkan bahwa penerapan algoritma Random Forest dengan seleksi fitur BestFirst sangat efektif dalam prediksi medis, khususnya kelangsungan hidup pasien gagal jantung. Pendekatan ini relevan untuk diterapkan

pada studi prediksi indikasi autisme, karena metode Random Forest mampu menangani data dengan banyak fitur dan memberikan akurasi tinggi dalam klasifikasi berbasis data medis.

2.2.3 Paper 3

Penelitian oleh Musyaffa et al. (2025) berfokus pada peningkatan performa prediksi ASD pada orang dewasa menggunakan algoritma Random Forest yang dikombinasikan dengan metode imputasi MissForest dan teknik oversampling SMOTE. Dataset yang digunakan adalah Autism Screening Adult dari UCI dengan jumlah data 704 sampel. Hasil penelitian menunjukkan bahwa penerapan SMOTE mampu mengatasi masalah ketidakseimbangan kelas dan meningkatkan akurasi model dari 70,17% menjadi 79,32%. Selain itu, nilai AUC-ROC meningkat signifikan dari 47,13% menjadi 85,84%. Studi ini menekankan pentingnya penanganan data yang hilang dan tidak seimbang dalam pengembangan sistem prediksi ASD yang lebih akurat dan dapat diandalkan

2.2.3.1 Tujuan Penelitian

Penelitian ini bertujuan untuk meningkatkan kinerja prediktif dalam diagnosis Autism Spectrum Disorder (ASD) pada orang dewasa melalui integrasi algoritma Random Forest, imputasi data MissForest, dan teknik oversampling SMOTE guna mengatasi masalah nilai hilang dan ketidakseimbangan kelas pada dataset.

2.2.3.2 Metodologi Yang Digunakan

Penelitian ini menggunakan Autism Screening Adult Dataset dari UCI yang berisi 704 data dengan 20 fitur. Metode imputasi MissForest digunakan untuk menangani nilai hilang, sementara SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas. Proses klasifikasi dilakukan dengan algoritma Random Forest, dan evaluasi model dilakukan menggunakan 10-fold cross validation dengan metrik akurasi, precision, recall, F1-score, dan AUC-ROC.

2.2.3.3 Temuan Utama

Hasil penelitian menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan performa model. Akurasi meningkat dari 70,17% menjadi 79,32%, precision dari 58,47% menjadi 79,55%, recall dari 70,17% menjadi 79,32%, F1-score dari 61,98% menjadi 79,28%, dan AUC-ROC dari 47,13% menjadi 85,84%. Ini membuktikan bahwa penanganan ketidakseimbangan data dan nilai hilang penting dalam meningkatkan kemampuan model prediksi.

2.2.3.4 Kesimpulan Penelitian

Penggunaan algoritma Random Forest dikombinasikan dengan imputasi MissForest dan penyeimbangan kelas menggunakan SMOTE secara efektif meningkatkan performa diagnosis prediktif terhadap ASD. Pendekatan ini sangat sesuai untuk digunakan pada studi prediksi autisme, termasuk pada peserta didik SLB, karena mampu menangani data tidak seimbang dan atribut yang hilang, yang sering terjadi dalam data medis dan pendidikan.