

BAB II TINJAUAN PUSTAKA

Penelitian mengenai klasifikasi *machine learning* di Indonesia, khususnya pada bidang medis dalam mengklasifikasi penyakit DBD sudah pernah dilakukan jauh sebelum penelitian ini disusun. Penelitian pertama dilakukan oleh Nurul Azam Mohd Salim dkk. yang membahas prediksi wabah demam berdarah *dengue* (DBD) di Selangor, Malaysia, menggunakan pendekatan *machine learning*. Data yang digunakan berupa jumlah kasus mingguan DBD dari tahun 2013 hingga 2017 di lima distrik, serta variabel iklim seperti suhu, kelembapan, curah hujan, dan kecepatan angin. Algoritma yang digunakan antara lain *Decision Tree* (CART), *Artificial Neural Network* (ANN), *Naïve Bayes*, dan *Support Vector Machine* (SVM) dengan kernel *linear*, polinomial, dan RBF. Hasil terbaik diperoleh dari SVM kernel *linear* dengan akurasi 70%, spesifisitas 95%, presisi 56%, dan sensitivitas yang meningkat dari 14% menjadi 63,5% setelah proses under-sampling. Penelitian ini menunjukkan bahwa SVM *linear* efektif dalam membangun sistem prediksi dini berbasis data iklim untuk mendukung tindakan preventif dan pengambilan keputusan oleh pihak kesehatan [5].

Penelitian kedua yang dilakukan oleh Hamdani dkk. bertujuan untuk mengembangkan metode klasifikasi penyakit demam berdarah *dengue* (DBD) berdasarkan gejala klinis pasien menggunakan pendekatan *machine learning*. Studi ini menggunakan lima algoritma klasifikasi, yaitu C.45, *Decision Tree* (DT), *K-Nearest Neighbor* (KNN), *Random Forest* (RF), dan *Support Vector Machine* (SVM), dengan data sebanyak 110 pasien dari Rumah Sakit Dirgahayu Samarinda yang mencakup 18 gejala klinis serta nilai trombosit dan hemoglobin. Evaluasi dilakukan menggunakan teknik *cross-validation* dengan k-fold 3, 5, dan 10 serta metrik akurasi, presisi, dan *recall*. Hasilnya, SVM dan *Random Forest* menunjukkan performa terbaik dengan nilai akurasi, presisi, dan *recall* mencapai 99,1% pada k-fold 10, sementara algoritma lainnya seperti C.45 dan KNN menunjukkan performa yang sedikit lebih rendah. Penelitian ini membuktikan bahwa algoritma SVM dan RF sangat efektif dalam mengklasifikasikan jenis penyakit *dengue*, serta menekankan pentingnya *machine learning* dalam mendukung sistem diagnosis klinis yang cepat dan akurat [6].

Penelitian ketiga yang dilakukan oleh Bilal dkk. bertujuan untuk mengembangkan sistem diagnosis dini penyakit demam berdarah *dengue* (DBD) berbasis data klinis menggunakan teknik *machine learning* yang efisien. Lima algoritma yang digunakan dalam penelitian ini adalah *K-Nearest Neighbor* (KNN), *Gradient Boosting Classifier* (GBC), *Extra Tree Classifier* (ETC), *eXtreme Gradient Boosting* (XGB), dan *LightGBM*, dengan validasi model menggunakan *Holdout* dan *10-Fold Cross-Validation*. Dataset sebanyak 6694 data pasien dengan 22 atribut klinis diperoleh dari *Epidemiological Monitoring Center*, Yaman, dan diproses melalui tahap penanganan *missing value*, normalisasi Z-score, seleksi fitur menggunakan *Extra Tree*, serta *balancing* data dengan *SMOTE+ENN*. Hasil pengujian menunjukkan bahwa model ETC memiliki performa terbaik dengan akurasi, *F1-score*, presisi, *recall*, dan AUC di atas 99% baik pada *Holdout* maupun *10-Fold Cross-Validation*, serta unggul dalam efisiensi waktu pemrosesan dibandingkan model lainnya. Penelitian ini membuktikan bahwa kombinasi *balancing* data dan seleksi fitur dapat meningkatkan akurasi prediksi secara signifikan, dengan ETC menjadi algoritma terbaik untuk sistem prediksi dini DBD berbasis data klinis [7].

Penelitian keempat dilakukan oleh Sheng-Wen Huang dkk. yang mengembangkan model prognostik untuk memprediksi tingkat keparahan penyakit *dengue* menggunakan pendekatan *machine learning*. Data yang digunakan berasal dari 798 pasien terkonfirmasi *dengue* di *National Cheng Kung University Hospital*, Taiwan, selama outbreak DENV-2 tahun 2015, yang mencakup informasi demografis serta hasil uji cepat laboratorium seperti NS1 antigen, anti-DENV IgM, dan IgG. Lima algoritma diuji, yaitu ANN, SVM, *Random Forest*, GBM, dan *Logistic Regression*, dengan evaluasi menggunakan *stratified 10-fold cross-validation*. Hasilnya menunjukkan bahwa model ANN memiliki performa terbaik dengan AUROC 0,8324 dan *balance accuracy* 75,23%, serta fitur usia >60 tahun dan NS1 positif menjadi faktor paling berpengaruh. Meskipun demikian, keterbatasan tetap ada karena model hanya dilatih dari satu lokasi dan satu serotipe, sehingga diperlukan validasi eksternal untuk memastikan generalisasi model [8].

Penelitian yang dilakukan oleh Ramtin Zargari Marandi dkk. bertujuan untuk mengembangkan model *machine learning* dalam memprediksi dini kebocoran plasma

pada pasien yang diduga terinfeksi *Dengue* berdasarkan data klinis dalam 96 jam pertama sejak demam muncul. Studi ini menggunakan data dari 877 pasien di Sri Lanka (60,3% terkonfirmasi *Dengue*) dan menghasilkan 4768 instansi data yang dianalisis menggunakan algoritma *Random Forest*, *LightGBM*, dan *ensemble average stacking* bernama DENV5F-AS. Fitur paling informatif yang dipilih menggunakan metode *Minimum Description Length* (MDL) meliputi jumlah limfosit, kadar hemoglobin, hematokrit, usia, dan kadar enzim AST. Model DENV5F-AS mencapai AUC sebesar 0,80, dengan PPV 76,9%, NPV 72,5%, spesifisitas 87,9%, dan sensitivitas 54,8%, serta menunjukkan performa yang konsisten di berbagai waktu observasi dan kelompok pasien. Studi ini menyimpulkan bahwa pendekatan *machine learning* berbasis data klinis sederhana dapat membantu pengambilan keputusan di fasilitas kesehatan terbatas, dengan keunggulan pada interpretabilitas, efisiensi sumber daya, dan potensi penggunaan offline di wilayah endemis *Dengue* [9].

Penelitian yang dilakukan oleh Era Alfi Zurroh bertujuan untuk mengklasifikasikan diagnosis penyakit Demam Berdarah *Dengue* (DBD) menggunakan metode *hybrid Naïve Bayes-K Nearest Neighbor* (KNN) agar memperoleh akurasi yang lebih baik dibanding metode tunggal. Studi ini menggunakan data sekunder dari 240 pasien hasil tes hematologi di Puskesmas Temindung, Samarinda, dengan variabel usia, hemoglobin, leukosit, trombosit, dan hematokrit. Proses seleksi fitur dilakukan menggunakan algoritma *Naïve Bayes* dengan threshold probabilitas minimum 0,30, dilanjutkan klasifikasi oleh KNN dengan nilai $K=3$ dan validasi menggunakan 10-fold *cross validation*. Hasil penelitian menunjukkan bahwa *hybrid Naïve Bayes-KNN* mampu mencapai akurasi 95,4%, sensitivitas 96,2%, dan spesifisitas 94,4%, lebih tinggi dibandingkan KNN biasa dan PCA-KNN yang masing-masing akurasinya 87,9% dan 88,3%. Penelitian ini menyimpulkan bahwa kombinasi metode *Naïve Bayes-KNN* efektif meningkatkan akurasi diagnosis DBD karena adanya seleksi fitur berbasis probabilitas yang lebih relevan dibandingkan hanya mereduksi dimensi seperti pada PCA [10].

Penelitian yang dilakukan oleh Lisa N. Nisa dkk. bertujuan untuk mendeteksi *Dengue Shock Syndrome* (DSS) pada pasien Demam Berdarah *Dengue* menggunakan algoritma *Random Forest*. Studi ini menggunakan 501 data pasien DBD dari Rumah

Sakit Umum Daerah dr. Saiful Anwar periode 2014-2022 yang terdiri dari 40 fitur gejala, kondisi fisik, dan hasil laboratorium. Untuk mengatasi masalah data *imbalance*, dilakukan random *oversampling* dan SMOTE, serta seleksi fitur dengan *Principal Component Analysis* (PCA) dan *Information Gain*. Model *Random Forest* dengan kombinasi PCA dan random *oversampling* menghasilkan akurasi terbaik sebesar 91,1%, dengan *precision* 92%, *recall* 84%, dan F1-score 88%. Penelitian ini menunjukkan bahwa metode *oversampling* sederhana lebih efektif dibanding SMOTE untuk dataset dengan pola yang tidak terlalu kompleks, serta pentingnya pemilihan fitur relevan untuk meningkatkan performa deteksi DSS[11].

Penelitian yang dilakukan oleh Fahmi dkk. bertujuan untuk meningkatkan akurasi klasifikasi multi-kelas pada diagnosis klinis pasien DBD yang terdiri dari *Dengue Fever* (DF), *Dengue Hemorrhagic Fever* (DHF), dan *Dengue Shock Syndrome* (DSS) dengan mengatasi permasalahan data *imbalance*. Studi ini menggunakan 14.044 data pasien dari Dinas Kesehatan Kota Semarang periode 2016-2019 dengan 16 fitur klinis, seperti trombosit, hematokrit, hemoglobin, dan faktor klinis lainnya. Delapan algoritma diuji, yaitu AdaBoost, *Neural Network*, KNN, *Decision Tree*, *Random Forest*, *Naïve Bayes*, SVM, dan *Logistic Regression*, dengan penerapan *class weighting*, *random resampling*, SMOTE, serta seleksi fitur menggunakan ReliefF. Hasil terbaik dicapai oleh AdaBoost yang dikombinasikan dengan *class weighting* dan random resampling dengan akurasi mencapai 87%, lebih tinggi dibanding dataset asli yang hanya 72%. Penelitian ini menyimpulkan bahwa kombinasi teknik *balancing* dan seleksi fitur efektif meningkatkan akurasi pada data diagnosis klinis DBD yang *imbalance* serta dapat diaplikasikan pada kasus penyakit lain yang serupa[12].

Penelitian yang dilakukan oleh Kusuma dkk. bertujuan untuk mengembangkan model *machine learning* yang mampu mengklasifikasikan tingkat keparahan penyakit DBD dengan mengatasi permasalahan data yang tidak seimbang. Studi ini menggunakan 4.373 data pasien DBD dari Dinas Kesehatan Kabupaten Malang dengan fitur seperti usia, jenis kelamin, jumlah trombosit, hematokrit, dan hasil tes laboratorium IGM, IGG, serta NS1. Penelitian ini menerapkan algoritma XGBoost dan SVM dengan dukungan teknik *balancing* data menggunakan SMOTE-ENN serta optimasi *hyperparameter* dengan *Particle Swarm Optimization* (PSO) dan

GridSearchCV. Hasilnya, kombinasi SMOTE-ENN, XGBoost, dan PSO berhasil mencapai akurasi sebesar 95% dan F1-score 94%, jauh lebih baik dibanding model tanpa *balancing* yang hanya memperoleh akurasi 76% dan F1-score 49%. Studi ini menyimpulkan bahwa pendekatan *hybrid* ini efektif dalam meningkatkan performa klasifikasi multi-kelas DBD dan menjadi solusi optimal dalam menghadapi tantangan data kesehatan yang tidak seimbang[13].

Penelitian yang dilakukan oleh Hawa dkk. bertujuan untuk memprediksi klasifikasi penyebaran Demam Berdarah *Dengue* (DBD) di Kota Bandung dengan menerapkan algoritma *Decision Tree* yang dikombinasikan dengan *time-based feature expansion* untuk menangkap pola spasial-temporal. Studi ini menggunakan data dari 30 kecamatan di Bandung selama tahun 2017-2021 dengan 13 fitur yang meliputi jumlah penduduk, iklim (curah hujan, suhu, kelembaban), tingkat pendidikan, dan golongan darah. Untuk mengatasi masalah data tidak seimbang, dilakukan *random oversampling*, serta seleksi fitur menggunakan *SelectKBest*. Model terbaik dalam penelitian ini adalah Model 2A dengan akurasi 98%, F1-score 96%, *precision* 92%, dan *recall* 98%, jauh lebih baik dibandingkan model tanpa *time expansion* yang hanya mencapai 87,72%. Penelitian ini menyimpulkan bahwa kombinasi *Decision Tree* dan *time-based feature expansion* efektif meningkatkan prediksi klasifikasi penyebaran DBD hingga empat tahun ke depan, sehingga dapat menjadi alat bantu dalam perencanaan mitigasi kesehatan masyarakat di wilayah endemis seperti Bandung[14].

Penelitian ini memiliki perbedaan signifikan dibandingkan dengan penelitian terdahulu, terutama pada penggunaan dataset dari *Kaggle* yang telah dimodifikasi menggunakan parameter yang diberikan oleh WHO menjadi tiga kelas yaitu Non-DBD, DF, dan DHF, sehingga klasifikasi menjadi lebih spesifik dibandingkan dengan penelitian sebelumnya yang umumnya hanya menggunakan dua kelas atau tanpa modifikasi data. Selain itu, dalam penelitian ini diterapkan teknik *oversampling* menggunakan ADASYN (*Adaptive Synthetic Sampling Approach*) untuk menyeimbangkan distribusi kelas pada dataset, yang jarang dilakukan pada studi sebelumnya. Proses pelatihan model dilakukan secara sistematis melalui Pipeline dan *GridSearchCV* dengan tiga algoritma berbeda yakni *Random Forest*, *Gradient Boosting*, dan *Support Vector Machine*, dimana setiap model diuji dalam dua skenario

berbeda berdasarkan *parameter grid* untuk feature selection menggunakan *SelectFromModel* dengan estimator *Random Forest*. Pendekatan ini memungkinkan analisis perbandingan akurasi dari setiap skenario, yang belum banyak dikaji secara terstruktur dalam penelitian-penelitian sebelumnya.





Tabel 2.1 Perbandingan Penelitian

No	Peneliti	Algoritma	Dataset	Cross-validation	Penyeimbangan data	Evaluasi model	Akurasi unggul
1.	Nurul Salim dkk.(2023)[5]	SVM (Linear, Poly, RBF), ANN, <i>Decision Tree</i> , <i>Naïve Bayes</i>	Data kasus mingguan DBD dan data iklim di Selangor, Malaysia	Split data: 70% training, 30% testing	Ya, menggunakan Random Undersampling	Akurasi, Sensitivitas, Spesifisitas, Presisi	SVM Linear: akurasi 70%; sensitivitas naik signifikan dari ~14 % ke ~63,5 % setelah <i>balancing</i> ; fitur utama: <i>week-of-the-year</i>
2.	Hamdani dkk. (2023)[6]	SVM, <i>Random Forest</i> , C.45, <i>Decision Tree</i> , KNN	Data klinis 110 pasien DBD dari RS Dirgahayu Samarinda;	Cross-validation k-fold 3, 5, 10.	Tidak disebutkan	Akurasi, <i>Precision</i> , <i>Recall</i>	SVM & RF: Akurasi tertinggi 99.1% (k=3 dan 10); KNN akurasi terendah 93.6%; kelas DSS hanya 9 data, risiko <i>imbalance</i> tinggi
3.	Bilal dkk.(2023)[7]	KNN, GBC, ETC, XGB, LightGBM	Dataset pasien DBD dari Yaman (6694 data, 22 atribut klinis)	Holdout (70:30), 10-fold cross-validation.	Ya, SMOTE + ENN Hybrid	Akurasi, <i>Precision</i> , <i>Recall</i> , F1-score, AUC	ETC: Akurasi 99.12% (Holdout), 99.03% (10-Fold); tanpa <i>balancing</i> : akurasi hanya ~85% (GBC terbaik di kondisi <i>imbalanced</i>)
4.	Sheng-Wen Huang dkk.(2020)[8]	ANN, SVM, <i>Random Forest</i> , <i>Gradient Boosting</i> , <i>Logistic Regression</i>	Data pasien <i>dengue</i> DENV-2 di Taiwan (798 pasien)	<i>Stratified</i> 10-fold cross-validation.	Tidak disebutkan.	AUROC, <i>Balance Accuracy</i>	ANN: AUROC 0.8324, <i>balance accuracy</i> 75.23%; faktor penting: usia >60, NS1 positif, kombinasi IgM+/IgG+ meningkatkan risiko
5.	Ramtin Marandi dkk.(2024)[9]	<i>Random Forest</i> , <i>LightGBM</i> , <i>Stacking Ensemble</i>	Dataset pasien <i>Dengue</i> (877 pasien, 4768 <i>instance</i>) dari Sri Lanka	<i>Split data</i> 70% training, 30% testing;	Tidak disebutkan	AUC, Sensitivitas, Spesifisitas, PPV, NPV, <i>Brier Score</i>	DENV5F-AS: AUC 0.80; sensitivitas 50-54.8%, spesifisitas 89.1%, PPV 76.9%, NPV 72.5%; prediksi

No	Peneliti	Algoritma	Dataset	Cross-validation	Penyeimbangan data	Evaluasi model	Akurasi unggul
		(DENV5F-AS)		<i>Nested 2x5-fold CV</i>			plasma leakage di dengue
6.	Era Alfi Zurroh (2021)[10]	<i>Hybrid Naïve Bayes - KNN, KNN, PCA-KNN</i>	240 data pasien tes hematologi dari Puskesmas Temindung Samarinda	10-fold cross validation	Tidak disebutkan	Akurasi, Sensitivitas, Spesifisitas	95,4% (<i>Hybrid Naïve Bayes-KNN</i>)
7.	Bagus Pramono Aji (2025)	<i>Random Forest, Gradient Boosting, Support Vector Machine (SVM)</i>	Dataset Hematologi (1.511 pasien) dari General Hospital Jamalpur, Bangladesh	5-fold cross validation	Ya, menggunakan ADASYN	Akurasi, Precision, Recall, F1-score, Confusion Matrix	-