JURNAL TEKNOLOGI INFORMASI DIGITAL (JTID)
INTEGRASI PUBLIKASI DIGITAL
Vol. 2 No. 2, Juli 2026 E-ISSN (xxx-xxx)
Website : https://jurnal.ipdig.id/index.php/jtid

Klasifikasi Diabetes pada Data Tidak Seimbang Menggunakan Algoritma *HistGradientBoosting* dengan *Class Weighting dan Hyperparameter Tuning*

Aditya Pratama Werdana[1*], Cindy Setyowati[2], Endah Kurnia Fitri[3]

[1,2,3]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa, Bekasi, Indonesia.

Dikirimkan: 31-05-2026
Diterbitkan: 05-06-2026

Keywords:
Diabetes;
HistGradientBoosting;
Hyperparameter Tuning;
Imbalanced Data;
Machine Learning.

E-mail Penulis korespondensi:
adityawerdana5@gmail.com

Abstrak. Ketidakseimbangan kelas pada data kesehatan menjadi tantangan dalam pengembangan model prediksi karena dapat menyebabkan model lebih dominan mengenali kelas mayoritas dibandingkan kelas minoritas. Penelitian ini bertujuan mengembangkan model prediksi diabetes menggunakan algoritma HistGradientBoosting pada dataset dengan distribusi kelas tidak seimbang. Dataset terdiri dari 100.000 data dengan atribut gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, blood glucose level, dan diabetes sebagai target klasifikasi. Tahapan penelitian meliputi eksplorasi data, penghapusan data duplikat, preprocessing fitur numerik dan kategorik, pembagian data training dan testing, penerapan class weight, serta hyperparameter tuning menggunakan GridSearch Cross Validation. Model baseline menghasilkan accuracy 89,59%, precision 45,58%, recall 92,75%, F1-score 61,12%, ROC-AUC 97,78%, dan PR-AUC 88,40%. Setelah optimasi, model terbaik memperoleh accuracy 89,86%, precision 46,23%, recall 91,86%, F1-score 61,51%, ROC-AUC 97,77%, dan PR-AUC 88,32%. Hasil tersebut menunjukkan bahwa HistGradientBoosting mampu mempertahankan recall tinggi dalam mendeteksi diabetes pada data tidak seimbang meskipun precision kelas positif masih perlu ditingkatkan.

Abstract. *Class imbalance in healthcare data presents a significant challenge in developing prediction models because it may cause models to be biased toward the majority class while reducing their ability to recognize minority cases. This study aims to develop a diabetes prediction model using the HistGradientBoosting algorithm on an imbalanced dataset. The dataset consisted of 100,000 records with attributes including gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, blood glucose level, and diabetes as the classification target. The research stages included data exploration, duplicate removal, numerical and categorical feature preprocessing, training–testing data splitting, class weight implementation, and hyperparameter tuning using GridSearch Cross Validation. The baseline model achieved an accuracy of 89.59%, precision of 45.58%, recall of 92.75%, F1-score of 61.12%, ROC-AUC of 97.78%, and PR-AUC of 88.40%. After optimization, the best model achieved an accuracy of 89.86%, precision of 46.23%, recall of 91.86%, F1-score of 61.51%, ROC-AUC of 97.77%, and PR-AUC of 88.32%. These results indicate that HistGradientBoosting is capable of maintaining high recall in detecting diabetes within imbalanced data, although precision for the positive class still requires improvement.*

1. Pendahuluan

Diabetes merupakan salah satu penyakit tidak menular yang menjadi perhatian global karena jumlah penderitanya terus mengalami peningkatan dari tahun ke tahun. Penyakit ini terjadi ketika tubuh tidak mampu memproduksi insulin dalam jumlah yang cukup atau tidak dapat menggunakan insulin secara efektif, sehingga menyebabkan peningkatan kadar glukosa dalam darah [1]. Kondisi tersebut dapat memicu berbagai komplikasi serius seperti gangguan jantung, kerusakan ginjal, gangguan saraf, hingga penurunan fungsi penglihatan apabila tidak ditangani secara tepat [2]. Tingginya angka kasus diabetes menunjukkan perlunya upaya deteksi dan penanganan sejak dini untuk meminimalkan risiko komplikasi serta meningkatkan kualitas hidup penderita.

Perkembangan teknologi informasi dan ketersediaan data kesehatan telah membuka peluang pemanfaatan kecerdasan buatan dalam bidang medis. Salah satu pendekatan yang banyak digunakan adalah *machine learning*, yaitu metode yang memungkinkan sistem mempelajari pola dari data untuk menghasilkan prediksi atau klasifikasi secara otomatis [3]. Dalam konteks kesehatan, *machine learning* dimanfaatkan untuk membantu proses diagnosis, penilaian risiko penyakit, serta pengambilan keputusan klinis berdasarkan pola yang ditemukan pada data pasien [4].

Prediksi penyakit diabetes menggunakan *machine learning* menjadi topik yang terus berkembang karena penyakit ini dipengaruhi oleh berbagai faktor, seperti usia, indeks massa tubuh, kadar glukosa darah, riwayat hipertensi, penyakit jantung, hingga gaya hidup seperti kebiasaan merokok [5]. Kompleksitas hubungan antarvariabel tersebut membuat pendekatan konvensional sering kali memiliki keterbatasan dalam mengidentifikasi risiko secara cepat dan akurat.

Penelitian mengenai prediksi diabetes telah banyak dilakukan menggunakan berbagai algoritma *machine learning* dan *deep learning.* Penelitian oleh VijiyaKumar dkk. menggunakan algoritma *Random Forest* untuk prediksi diabetes dan menunjukkan bahwa metode *ensemble learning* tersebut mampu memberikan prediksi secara efektif dan cepat dibandingkan beberapa algoritma lain [6]. Penelitian lain oleh Jose Julian Hidayat, dkk. memanfaatkan *Deep Neural Network (DNN)* dengan penyesuaian hiperparameter berbasis *Bayesian Optimization* pada 15.000 data pasien dan menghasilkan performa tinggi dengan *accuracy* 97,14% serta AUC 0,9764, yang menegaskan pentingnya optimasi parameter dalam meningkatkan stabilitas model [7]. Selanjutnya, penelitian yang dilakukan oleh Hovi dkk. menerapkan *Support Vector Machine (SVM)* berbasis *Radial Basis Function (RBF)* dengan *Forward Selection* dan memperoleh akurasi sebesar 91%, sehingga menunjukkan bahwa *SVM-RBF* dapat menjadi alternatif yang baik dalam klasifikasi diabetes [8]. Selain itu, Jose Julian Hidayat, Daffa Eka Sujianto, Muhammad Randika Saputra, Erik Ahmad Ramdhani, Muhamad Jihansyah dan Yogi Nandya. mengembangkan klasifikasi diabetes menggunakan *Multi-Layer Perceptron (MLP)* berbasis jaringan syaraf tiruan pada dataset 100.000 data dan memperoleh *accuracy* 97,15% dengan *ROC-AUC* 0,9749 [4]. Berdasarkan beberapa penelitian tersebut, dapat diketahui bahwa prediksi diabetes terus berkembang melalui pemanfaatan berbagai algoritma, namun penggunaan *HistGradientBoosting* pada data tidak seimbang dengan pendekatan *hyperparameter* tuning masih relatif terbatas sehingga menjadi fokus dalam penelitian ini [9]. Pada penelitian lain, Hidayat dkk. mengevaluasi beberapa algoritma *ensemble learning* berbasis *boosting*, yaitu *Gradient Boosting*, *XGBoost*, dan *CatBoost*, dengan hasil *accuracy* mencapai 97%, namun nilai *recall* pada kelas diabetes masih *relatif* rendah yaitu sekitar 0,69 sehingga menunjukkan adanya tantangan dalam mendeteksi seluruh kasus positif [10]. Meskipun memiliki potensi yang baik, penerapan *machine learning* pada data kesehatan masih menghadapi sejumlah tantangan [11]. Salah satu tantangan utama adalah ketidakseimbangan kelas atau *imbalanced data*, yaitu kondisi ketika jumlah data pada satu kelas jauh lebih dominan dibandingkan kelas lainnya [12]. Pada kasus prediksi diabetes, data non-diabetes umumnya lebih banyak dibandingkan data diabetes [13]. Kondisi ini dapat menyebabkan model cenderung memprediksi kelas mayoritas dan mengurangi kemampuan dalam mengenali pasien yang benar-benar memiliki risiko diabetes [14]. Oleh sebab itu, diperlukan strategi yang mampu membantu model memberikan perhatian yang lebih seimbang terhadap seluruh kelas data. Permasalahan ketidakseimbangan data *(imbalanced data)* menjadi tantangan penting dalam klasifikasi karena dapat menyebabkan model cenderung memprediksi kelas mayoritas dan menurunkan kemampuan deteksi pada kelas minoritas. Berbagai penelitian telah mengembangkan metode penanganan ketidakseimbangan data, terutama melalui pendekatan *oversampling*. Arifiyanti dan Wahyuni menunjukkan bahwa metode *SMOTE* mampu meningkatkan performa beberapa model klasifikasi seperti *Logistic Regression, KNN,* dan *Naïve Bayes*, meskipun tidak memberikan perubahan signifikan pada *Decision Tree* [15]. Pendekatan serupa juga diterapkan pada berbagai domain tidak hanya dalam bidang kesehatan atau dalam prediksi diabetes saja, tetapi seperti prediksi kebangkrutan perusahaan menggunakan *Random Forest* yang mengalami peningkatan performa sebesar 7,40% setelah penerapan *SMOTE* [16]. Serta deteksi transaksi mencurigakan yang menunjukkan kombinasi *Decision Tree* dan *SMOTE* menghasilkan *F1-score* 0,85 serta AUC 0,95 [17]. Pada bidang kesehatan, Zaenur Rozikin dan Jose Julian Hidayat membandingkan *SMOTE* dan *ADASYN* pada klasifikasi diabetes menggunakan *CatBoost* dan menemukan bahwa *SMOTE* memberikan peningkatan *recall* serta *F1-score*

yang lebih optimal dibandingkan *ADASYN* [18]. Penelitian lain oleh Alwaliyanto dkk. juga membuktikan bahwa *ADASYN* mampu meningkatkan sensitivitas model *SVM* pada klasifikasi stroke melalui peningkatan *recall* dan *F1-score* [19]. Selain itu, penerapan *SMOTE* pada klasifikasi gangguan kecemasan mahasiswa meningkatkan recall dari 0,41 menjadi 0,69 meskipun terjadi sedikit penurunan akurasi [20]. Studi komparatif Nugroho dkk. menunjukkan bahwa penggunaan *SMOTE* pada berbagai algoritma klasifikasi mampu meningkatkan keseimbangan performa model, terutama pada *Random Forest* dan *Logistic Regression* [21]. Pada penelitian internasional, Wang dkk. mengombinasikan *SMOTE*, *Borderline-SMOTE,* dan *SMOTE-ENN* pada *Bayesian Network* untuk prediksi diabetes dan menemukan bahwa *SMOTE-ENN* menghasilkan peningkatan performa klasifikasi paling signifikan [22]. Namun demikian, penelitian Yang dkk. menunjukkan bahwa random *oversampling* dan *undersampling* tidak selalu meningkatkan *AUROC* pada data kesehatan observasional sehingga pemilihan teknik penanganan ketidakseimbangan perlu disesuaikan dengan karakteristik dataset dan tujuan evaluasi model [23].

Berdasarkan berbagai penelitian terdahulu, prediksi diabetes telah banyak dikembangkan menggunakan algoritma seperti *Random Forest, Support Vector Machine (SVM), Deep Neural Network (DNN), Multi-Layer Perceptron (MLP),* serta beberapa metode *ensemble berbasis boosting* seperti *Gradient Boosting, XGBoost,* dan *CatBoost.* Penelitian mengenai penanganan ketidakseimbangan data juga telah menunjukkan bahwa teknik *oversampling* seperti *SMOTE* dan *ADASYN* mampu meningkatkan sensitivitas model terhadap kelas minoritas. Namun, sebagian besar penelitian masih berfokus pada algoritma *boosting* konvensional atau pendekatan *oversampling*, sementara penggunaan *HistGradientBoosting* sebagai metode *boosting* berbasis *histogram* pada kasus klasifikasi diabetes dengan data tidak seimbang masih relatif terbatas. Selain itu, belum banyak penelitian yang mengombinasikan *HistGradientBoosting* dengan strategi penyesuaian bobot kelas (*class weight*) dan *hyperparameter tuning* untuk meningkatkan kemampuan model dalam mendeteksi kelas diabetes tanpa bergantung pada proses *oversampling*. Oleh karena itu, *novelty* pada penelitian ini terletak pada penerapan dan optimasi algoritma *HistGradientBoosting* menggunakan *hyperparameter tuning* pada *dataset* diabetes tidak seimbang dengan pendekatan *class weight*, sehingga diharapkan mampu memberikan model prediksi yang efisien, adaptif, dan memiliki sensitivitas yang baik terhadap kelas minoritas.

Berdasarkan permasalahan tersebut, penelitian ini berfokus pada pengembangan model prediksi penyakit diabetes menggunakan algoritma *HistGradientBoosting* pada data yang tidak seimbang. Penelitian dilakukan melalui tahapan pengolahan data, *preprocessing* fitur, penanganan ketidakseimbangan kelas, serta optimasi model menggunakan *hyperparameter tuning*. Pendekatan ini diharapkan dapat menghasilkan model yang lebih adaptif dalam mengenali pola diabetes sehingga berpotensi mendukung pengembangan sistem prediksi berbasis data sebagai alat bantu dalam deteksi dini penyakit diabetes.

2. Metode Penelitian

2.1. Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode *machine learning* untuk membangun model prediksi penyakit diabetes [24]. Pendekatan kuantitatif dipilih karena penelitian berfokus pada pengolahan data numerik, pengujian model secara komputasional, serta pengukuran performa berdasarkan metrik evaluasi statistik [25]. Algoritma yang digunakan adalah *HistGradientBoosting* karena memiliki kemampuan membangun model klasifikasi berbasis *boosting* dengan efisiensi tinggi pada dataset berukuran besar serta mampu mempelajari hubungan kompleks antarfitur kesehatan [26].

2.2. *Dataset* dan Pengumpulan Data

Dataset yang digunakan dalam penelitian ini merupakan dataset diabetes yang terdiri dari 100.000 data dengan sembilan atribut, yaitu *gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, blood glucose level,* dan *diabetes* sebagai target klasifikasi. Data diperoleh dari sumber dataset terbuka yang umum digunakan dalam penelitian *machine learning* kesehatan yang dapat diakses melalui tautan *https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset*. *Dataset* ini dipilih karena memiliki jumlah data yang besar serta memuat berbagai faktor klinis dan gaya hidup yang relevan untuk mendukung proses prediksi penyakit diabetes.

2.3. Pra-pemrosesan Data

Tahap *preprocessing* dilakukan untuk memastikan kualitas data sebelum digunakan dalam proses pemodelan [27]. Proses ini meliputi pemeriksaan *missing value,* analisis statistik deskriptif, penghapusan data duplikat, serta identifikasi tipe data numerik dan kategorik [28]. Fitur kategorik seperti *gender* dan *smoking history* ditransformasikan menggunakan teknik *encoding*, sedangkan fitur numerik dipersiapkan agar dapat diproses

secara optimal oleh model *machine learning* [29]. Setelah *preprocessing* selesai, data dibagi menjadi *data training* dan *testing* guna mendukung proses pelatihan dan evaluasi model [30].

2.4. Penanganan Ketidakseimbangan

Ketidakseimbangan data menjadi salah satu tantangan dalam klasifikasi diabetes karena jumlah data non-diabetes lebih dominan dibandingkan data diabetes. Kondisi ini dapat menyebabkan model memiliki kecenderungan memprediksi kelas mayoritas dan menurunkan kemampuan deteksi pada kelas minoritas. Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan pendekatan *class weight* yang memberikan bobot lebih besar pada kelas minoritas sehingga model dapat mempelajari pola diabetes secara lebih proporsional tanpa melakukan perubahan distribusi data melalui teknik *oversampling* atau *undersampling* [31],[32].

2.5. Pemodelan *HistGradientBoosting*

Pemodelan dilakukan menggunakan algoritma *HistGradientBoosting* yang termasuk dalam metode *ensemble learning* berbasis *boosting*. Algoritma ini bekerja dengan membangun sejumlah *weak learner* secara bertahap, di mana setiap iterasi berupaya memperbaiki kesalahan prediksi model sebelumnya [33]. Pendekatan *histogram-based learning* pada *HistGradientBoosting* memungkinkan proses pelatihan menjadi lebih cepat dan efisien, terutama pada dataset berukuran besar [34]. Pada penelitian ini, model awal atau baseline dibangun terlebih dahulu untuk memperoleh gambaran performa awal sebelum dilakukan optimasi.

2.5. Tahap *Hyperparameter Tuning*

Setelah model baseline dibangun, tahap berikutnya adalah melakukan *hyperparameter tuning* untuk memperoleh konfigurasi model yang lebih optimal. Proses *tuning* dilakukan menggunakan *GridSearch Cross Validation* dengan menguji berbagai kombinasi parameter seperti learning *rate*, *max depth, max iteration*, *minimum samples leaf*, dan *regularization* [35]. Pendekatan ini bertujuan untuk menemukan parameter terbaik yang mampu meningkatkan keseimbangan performa model serta mengurangi risiko *overfitting* maupun *underfitting* selama proses pembelajaran [36].

2.6. Evaluasi Model

Evaluasi model dilakukan untuk mengukur kemampuan *HistGradientBoosting* dalam memprediksi penyakit diabetes secara menyeluruh. Pengujian dilakukan menggunakan *confusion matrix* serta beberapa metrik evaluasi yaitu *accuracy, precision, recall, F1-score, ROC-AUC*, dan *PR-AUC* [37]. *Accuracy* digunakan untuk mengukur tingkat ketepatan prediksi secara umum [38], sedangkan *precision*, *recall*, dan *F1-score* digunakan untuk mengevaluasi performa model terhadap kelas diabetes sebagai kelas minoritas [39],[40]. Selain itu, *ROC-AUC* dan *PR-AUC* digunakan untuk menilai kemampuan model dalam membedakan kelas serta mengevaluasi kualitas prediksi pada data tidak seimbang [41].

3. Hasil dan Pembahasan

Penelitian ini menggunakan dataset diabetes yang terdiri dari 100.000 data dengan sembilan atribut, yaitu *gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, blood glucose level,* dan *diabetes* sebagai target klasifikasi seperti pada Tabel 1.

Tabel 1. Dataset Penelitian

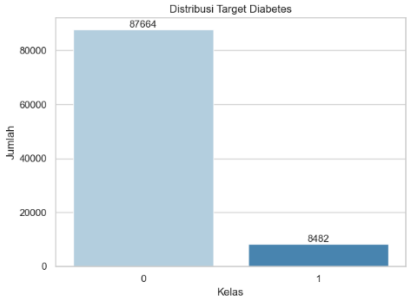
No	Gender	Age	Hypertension	Heart Disease	Smoking History	BMI	HbA1c Level	Blood Glucose Level	Diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0
10	Female	53.0	0	0	never	27.32	6.1	85	0
11	Female	54.0	0	0	former	54.70	6.0	100	0
12	Female	78.0	0	0	former	36.05	5.0	130	0
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
96144	Female	67.0	0	0	never	25.69	5.8	200	0
96145	Female	76.0	0	0	No Info	27.32	5.0	160	0

Hasil eksplorasi data pada Tabel 2 menunjukkan bahwa dataset tidak memiliki *missing value,* namun ditemukan sebanyak 3.854 data duplikat sehingga dilakukan proses pembersihan data dan jumlah data menjadi 96.146 *record.*

Tabel 2. Missing Value Table

Atribut	Missing Value
gender	0
age	0
hypertension	0
heart_disease	0
smoking_history	0
bmi	0
HbA1c_level	0
blood_glucose_level	0
diabetes	0

Distribusi kelas menunjukkan adanya ketidakseimbangan data, di mana kelas non-diabetes mendominasi sebesar 91,18% sedangkan kelas diabetes hanya sebesar 8,82%. Kondisi ini menunjukkan bahwa dataset termasuk kategori imbalanced data yang berpotensi memengaruhi performa model dalam mendeteksi kelas minoritas.



Gambar 1. Distribusi Target

Tahap *preprocessing* dilakukan melalui identifikasi fitur numerik dan kategorik, transformasi data, serta pembagian dataset menjadi *data training* dan *testing*. *Data training* terdiri dari 76.916 data sedangkan *data testing* berjumlah 19.230 data.

Tabel 3. Distribusi *Data Training* dan *Data Testing*

Dataset	Shape	Jumlah Data	Jumlah Fitur
X_train	(76916, 8)	76.916	8
X_test	(19230, 8)	19.230	8
y_train	(76916,)	76.916	-
y_test	(19230,)	19.230	-

Untuk mengatasi ketidakseimbangan kelas diterapkan pendekatan *class weight*, sehingga kelas diabetes memperoleh bobot lebih besar dibandingkan kelas non-diabetes. Pendekatan ini bertujuan agar model dapat memberikan perhatian yang lebih proporsional terhadap data diabetes selama proses pembelajaran tanpa mengubah distribusi asli dataset.

Tabel 4. Hasil Pembobotan

Kelas	**Bobot (Weight)**
Class 0 (Non-Diabetes)	0.5484
Class 1 (Diabetes)	5.6673

Hasil *classification report* menunjukkan bahwa model *HistGradientBoosting* memiliki performa yang berbeda pada masing-masing kelas. Pada kelas non-diabetes (kelas 0), model memperoleh *precision* sebesar 0,99, *recall* 0,89, dan *F1-score* 0,94, yang menunjukkan kemampuan sangat baik dalam mengidentifikasi pasien non-diabetes. Sementara itu, pada kelas diabetes (kelas 1), model menghasilkan *precision* sebesar 0,46, *recall* sebesar 0,93, dan *F1-score* sebesar 0,61. Nilai *recall* yang tinggi menunjukkan bahwa sebagian besar kasus diabetes berhasil terdeteksi oleh model, yang penting dalam konteks deteksi dini penyakit untuk meminimalkan kasus *false negative*. Namun, *precision* yang relatif rendah menunjukkan masih terdapat cukup banyak prediksi positif yang sebenarnya bukan diabetes *(false positive).* Nilai *macro average recall* sebesar 0,91 menunjukkan bahwa model memiliki sensitivitas yang baik pada kedua kelas, sedangkan *weighted average* yang tinggi dipengaruhi oleh dominasi jumlah data pada kelas non-diabetes. Secara keseluruhan, model *HistGradientBoosting* menunjukkan kemampuan

yang baik dalam mendeteksi kasus diabetes pada data tidak seimbang, meskipun peningkatan *precision* masih diperlukan agar prediksi positif menjadi lebih akurat.

Tabel 5. *Classification Report* dari *HistGradientBoosting Baseline*

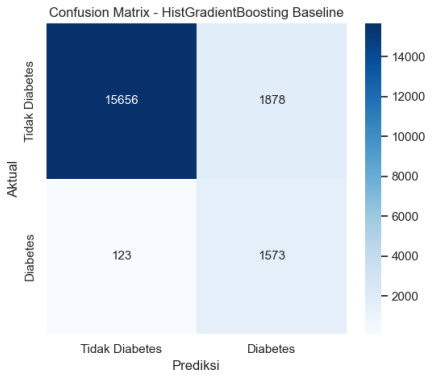
Class	Precision	Recall	F1-Score	Support
0 (Non-Diabetes)	0.99	0.89	0.94	17.534
1 (Diabetes)	0.46	0.93	0.61	1.696
Accuracy	-	-	0.90	19.230
Macro Avg	0.72	0.91	0.78	19.230
Weighted Avg	0.94	0.90	0.91	19.230

Model *HistGradientBoosting* hasil *hyperparameter tuning* menunjukkan performa sangat baik pada kelas non-diabetes dengan *precision* 0,99 dan *F1-score* 0,94. Pada kelas diabetes, model memperoleh *recall* tinggi sebesar 0,92 yang menunjukkan sebagian besar kasus diabetes berhasil terdeteksi. Namun *precision* sebesar 0,46 menunjukkan masih terdapat *false positive* yang cukup tinggi sehingga tidak semua prediksi diabetes benar-benar positif. Nilai *weighted average* sebesar 0,91 dan *accuracy* 0,90 memperlihatkan performa keseluruhan model yang stabil, sementara *macro average F1-score* sebesar 0,78 menunjukkan model cukup baik dalam menangani data tidak seimbang meskipun peningkatan precision pada kelas diabetes masih menjadi tantangan.

Tabel 6. Classification Report dari HistGradientBoosting setelah Class Weighting dan Hyperparameter Tuning

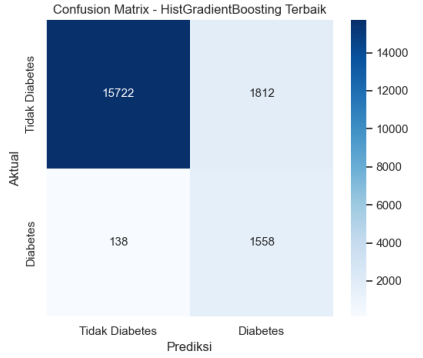
Class	Precision	Recall	F1-Score	Support
0 (Non-Diabetes)	0.99	0.90	0.94	17.534
1 (Diabetes)	0.46	0.92	0.62	1.696
Accuracy	-	-	0.90	19.230
Macro Average	0.73	0.91	0.78	19.230
Weighted Average	0.94	0.90	0.91	19.230

Pada gambar 2 model *baseline*, sebanyak 1.573 kasus diabetes berhasil diprediksi dengan benar (*TP*) dan hanya 123 kasus diabetes yang gagal terdeteksi (*FN*), menunjukkan sensitivitas model yang tinggi.



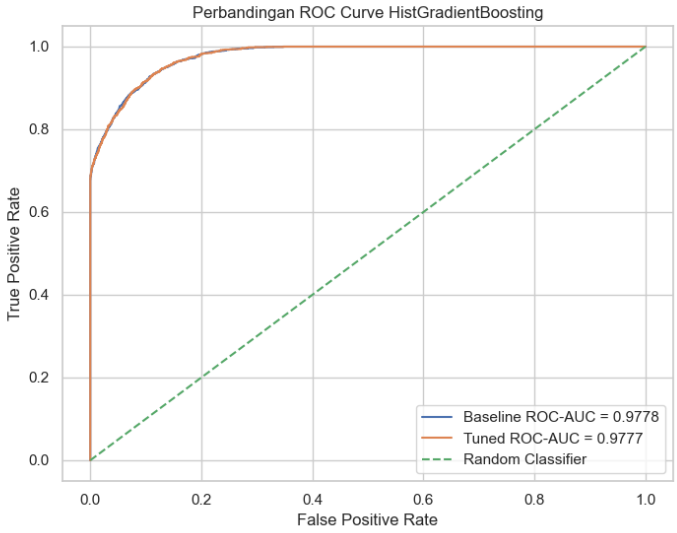
Gambar 2. *Confusion Matrix* dari *HistGradientBoosting Baseline*

Pada gambar 3 setelah *hyperparameter tuning,* jumlah prediksi benar pada kelas non-diabetes meningkat dari 15.656 menjadi 15.722 dan *false positive* menurun dari 1.878 menjadi 1.812. Namun, jumlah *false negative* meningkat dari 123 menjadi 138 sehingga *true positive* sedikit menurun dari 1.573 menjadi 1.558.



Gambar 3. Confusion Matrix dari HistGradientBoosting setelah Class Weighting dan Hyperparameter Tuning

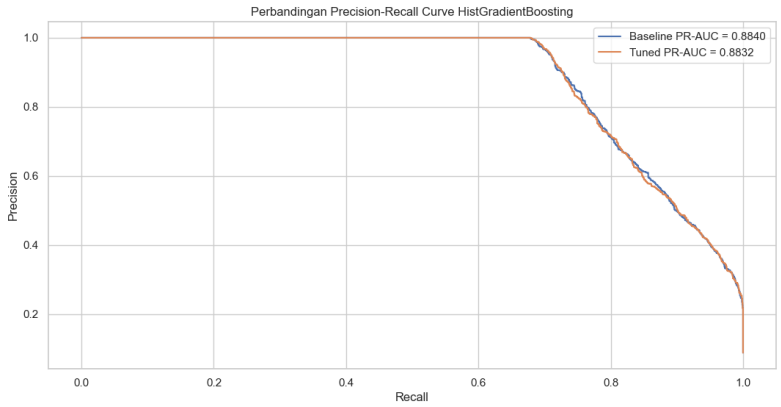
Pada grafik ROC di Gambar 4 terlihat bahwa kurva model *baseline* dan *tuned* hampir saling bertumpuk. Nilai *ROC-AUC baseline* sebesar 0,9778, sedangkan *ROC-AUC model tuned* sebesar 0,9777. Selisih yang sangat kecil (0,0001) menunjukkan bahwa kemampuan model dalam membedakan kelas positif dan negatif tetap sangat baik setelah *tuning*, namun tidak mengalami peningkatan besar. Nilai *ROC-AUC* yang mendekati 1 menunjukkan bahwa model memiliki kemampuan diskriminasi yang sangat tinggi dan jauh lebih baik dibandingkan *classifier* acak yang direpresentasikan oleh garis diagonal.



Gambar 4. *ROC-Curve* dari Perbandingan Model Sebelum dan Sesudah *Improvment*

Gambar 4 Kurva ROC pada model *HistGradientBoosting baseline* menunjukkan nilai *ROC-AUC* sebesar 0,9778, yang mengindikasikan kemampuan diskriminasi model sangat baik dalam membedakan pasien diabetes dan non-diabetes. Kurva *ROC* terlihat mendekati sudut kiri atas dan berada jauh di atas garis *diagonal random classifier*, menandakan bahwa model memiliki tingkat *true positive rate* yang tinggi dengan *false positive rate* yang relatif rendah. Nilai *ROC-AUC* yang mendekati 1 menunjukkan bahwa model mampu melakukan klasifikasi secara efektif meskipun dataset memiliki distribusi kelas yang tidak seimbang.

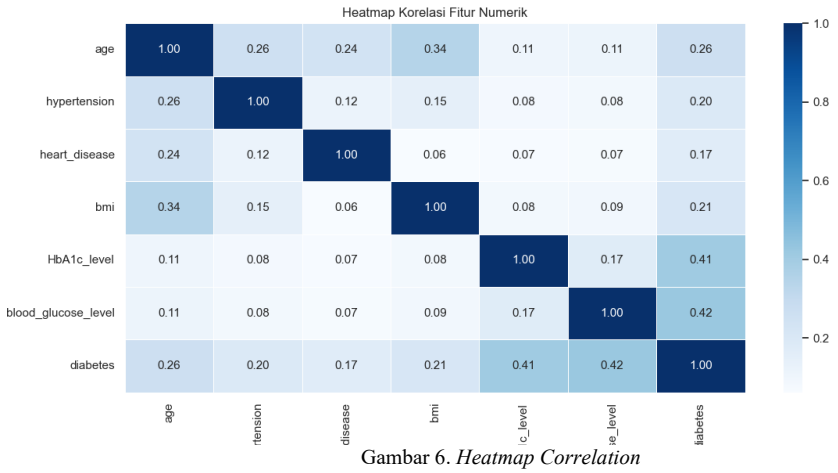
Grafik Precision-Recall pada gambar 5 juga menunjukkan pola yang hampir identik antara kedua model. Nilai *PR-AUC baseline* sebesar 0,8840, sedangkan *PR-AUC model tuned* sebesar 0,8832. Penurunan yang sangat kecil ini mengindikasikan bahwa *tuning hyperparameter* tidak meningkatkan keseimbangan antara *precision* dan *recall*. Kurva yang berada pada area atas grafik menunjukkan bahwa model masih mampu mempertahankan *precision* yang tinggi pada berbagai tingkat *recall*, sehingga performanya tetap baik terutama dalam mendeteksi kelas positif.



Gambar 5. *Precision-Recall Curve* dari *HGB Baseline*

Berdasarkan *heatmap* korelasi pada gambar 6, hubungan antar fitur numerik pada dataset diabetes umumnya berada pada kategori rendah hingga sedang, yang ditunjukkan oleh nilai koefisien korelasi yang sebagian besar

berada di bawah 0,5. Kondisi ini mengindikasikan bahwa tidak terdapat masalah multikolinearitas yang kuat antar variabel independen, sehingga setiap fitur masih berpotensi memberikan informasi yang berbeda dalam proses prediksi diabetes.



Gambar 6. *Heatmap Correlation*

Variabel yang memiliki korelasi paling tinggi dengan target diabetes adalah *blood_glucose_level* dengan nilai korelasi sebesar 0,42, diikuti oleh *HbA1c_level* sebesar 0,41. Hasil ini menunjukkan bahwa peningkatan kadar glukosa darah dan *HbA1c* cenderung diikuti oleh meningkatnya kemungkinan seseorang menderita diabetes. Temuan tersebut sesuai dengan pengetahuan medis karena kedua variabel tersebut merupakan indikator utama yang digunakan dalam diagnosis diabetes.

Selain itu, fitur *age* memiliki korelasi sebesar 0,26 terhadap diabetes, sedangkan *BMI* memiliki korelasi sebesar 0,21 dan *hypertension* sebesar 0,20. Nilai korelasi positif ini menunjukkan bahwa bertambahnya usia, meningkatnya indeks massa tubuh, serta adanya riwayat hipertensi berhubungan dengan peningkatan risiko diabetes. Sementara itu, *heart_disease* memiliki korelasi paling rendah terhadap diabetes, yaitu sebesar 0,17, yang menunjukkan hubungan yang relatif lemah dibandingkan fitur lainnya.

Di antara fitur independen, korelasi tertinggi ditemukan antara *age* dan *BMI* sebesar 0,34, yang menunjukkan bahwa peningkatan usia cenderung diikuti oleh peningkatan *BMI*. Namun, nilai tersebut masih tergolong rendah hingga sedang sehingga tidak menimbulkan masalah redundansi fitur yang signifikan. Korelasi antara *HbA1c_level* dan *blood_glucose_level* sebesar 0,17 juga menunjukkan adanya hubungan positif, tetapi tidak cukup kuat untuk menyebabkan ketergantungan yang tinggi antar variabel.

Secara keseluruhan, hasil analisis korelasi menunjukkan bahwa *blood_glucose_level* dan *HbA1c_level* merupakan fitur yang paling berpengaruh terhadap status diabetes, sedangkan fitur lainnya memberikan kontribusi tambahan dengan tingkat hubungan yang lebih rendah. Tidak adanya korelasi yang sangat tinggi antar fitur independen mengindikasikan bahwa seluruh variabel masih layak dipertahankan dalam proses pembangunan model klasifikasi diabetes.

4. Kesimpulan

Penelitian ini berhasil membangun model prediksi penyakit diabetes menggunakan algoritma *HistGradientBoosting* pada dataset dengan distribusi kelas tidak seimbang. Tahapan penelitian meliputi *preprocessing* data, penghapusan data duplikat, pembagian *data training* dan *testing*, penerapan *class weight*, serta *hyperparameter tuning* menggunakan *GridSearch Cross Validation*. Hasil evaluasi menunjukkan bahwa model *baseline* memperoleh *accuracy* sebesar 89,59%, *precision* 45,58%, *recall* 92,75%, *F1-score* 61,12%, *ROC-AUC* 97,78%, dan *PR-AUC* 88,40%. Setelah dilakukan optimasi, model terbaik menghasilkan *accuracy* 89,86%, *precision* 46,23%, *recall* 91,86%, *F1-score* 61,51%, *ROC-AUC* 97,77%, dan *PR-AUC* 88,32%. Hasil tersebut menunjukkan bahwa *HistGradientBoosting* mampu mempertahankan kemampuan deteksi diabetes yang tinggi pada data tidak seimbang, terutama ditunjukkan oleh nilai *recall* dan *ROC-AUC* yang tinggi. Meskipun peningkatan performa setelah tuning relatif kecil, optimasi parameter berhasil menghasilkan model yang lebih seimbang dan stabil dalam proses klasifikasi.

5. Saran

Penelitian selanjutnya disarankan untuk mengeksplorasi teknik penanganan ketidakseimbangan data seperti *SMOTE, ADASYN*, atau *hybrid sampling* guna meningkatkan *precision* pada kelas diabetes tanpa menurunkan *recall secara* signifikan. Selain itu, perbandingan dengan algoritma ensemble lain seperti *GBM, LightGBM,* dan *CatBoost* dapat dilakukan untuk memperoleh model dengan performa yang lebih optimal. Penggunaan *threshold tuning* dan teknik *probability calibration* juga dapat dipertimbangkan agar prediksi model menjadi lebih presisi. Dari sisi data, penambahan atribut klinis maupun riwayat kesehatan lain berpotensi meningkatkan kemampuan model dalam mengenali pola diabetes secara lebih komprehensif sehingga dapat mendukung pengembangan sistem prediksi risiko diabetes yang lebih akurat dan aplikatif.

Daftar Rujukan

[1] H. Husain, S. Ramadani, and N. Magfirah Ilyas, "Literature Review: Analisis Faktor Penyebab Penyakit Degeneratif (Diabetes Mellitus) pada Metabolisme Karbohidrat," *Indones. J. Sci. Public Heal.*, vol. 2, no. 2 SE-Articles, pp. 258–270, Sep. 2025, [Online]. Available: https://yici-journal.id/ijsph/article/view/29

[2] N. N. Rosyidah and E. A. Cahyono, "DIABETES MELITUS TIPE 2 ; ARTIKEL REVIEW," *Enferm. Cienc.*, vol. 3, no. 1, pp. 44–63, Feb. 2025, doi: 10.56586/ec.v3i1.74.

[3] D. Fauzi, M. Isnan, K. S. Nugroho, and B. Pardamean, "Machine Learning & Deep Learning".

[4] I. A. Wibowo and M. Kom, *KECERDASAN BISNIS (BI) DIDUKUNG AI: Meningkatkan prediksi dan pembuatan keputusan dengan ML (Machine Learning)*. yayasan penerbit, 2025.

[5] D. Rahmawati, "Kualitas Hidup Pasien Diabetes Melitus dan Hipertensi dalam Program Penyakit Kronis (Prolanis) di Indonesia: Narative Review," *J. Mandala Pharmacon Indones.*, vol. 10, no. 1 SE-Review Article, pp. 116–122, Jun. 2024, doi: 10.35311/jmpi.v10i1.531.

[6] K. VijiyaKumar, B. Lavanya, I. Nirmala, and S. S. Caroline, "Random Forest Algorithm for the Prediction of Diabetes," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–5. doi: 10.1109/ICSCAN.2019.8878802.

[7] J. J. Hidayat *et al.*, "Prediksi Diabetes Menggunakan Deep Neural Network dengan Penyesuaian Hiperparameter Berbasis Bayesian Optimization," *J. Pract. Comput. Sci.*, vol. 5, no. 2, pp. 130–143, Jan. 2026, doi: 10.37366/jpcs.v5i2.6419.

[8] H. S. W. Hovi, A. Id Hadiana, and F. Rakhmat Umbara, "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)," *Informatics Digit. Expert*, vol. 4, no. 1 SE-Articles, pp. 40–45, Jan. 2023, doi: 10.36423/index.v4i1.895.

[9] J. J. Hidayat, D. E. Sujianto, M. R. Saputra, E. A. Ramdhani, M. Jihansyah, and Y. Nandya, "Klasifikasi Penyakit Diabetes Melitus Berbasis Jaringan Syaraf Tiruan Menggunakan Algoritma Multi-Layer Perceptron," *J. Komput. Teknol. Inf. Sist. Komput.*, vol. 5, no. 1, pp. 401–411, May 2026, doi: 10.62712/juktisi.v5i1.1042.

[10] J. J. Hidayat, M. R. Saputra, A. R. Sigand, A. L. N. Fadilah, M. D. I. Amin, and A. R. Ramadhan, "Evaluasi Kinerja Algoritma Ensemble Learning Pada Klasifikasi Penyakit Diabetes Berbasis Boosting Method," *J. Surya Inform.*, vol. 16, no. 1 SE-Articles, pp. 71–80, May 2026, doi: 10.48144/suryainformatika.v16i1.2424.

[11] J. J. Hidayat, F. F. Azhari, T. M. Husna, A. N. Fahmayani, N. N. Pradana, and C. Setyowati, "Perbandingan Kinerja Algoritma Machine Learning Dalam Prediksi Kesehatan Mental Dan Burnout Mahasiswa," *J. Surya Inform.*, vol. 16, no. 1 SE-Articles, pp. 32–42, May 2026, doi: 10.48144/suryainformatika.v16i1.2420.

[12] Dina Wulan Yekti rahayu, Khothibul Umam, and Maya Rini Handayani, "Performance of Machine Learning Algorithms on Imbalanced Sentiment Datasets Without Balancing Techniques," *J. Appl. Informatics Comput.*, vol. 9, no. 3 SE-Articles, pp. 998–1005, Jun. 2025, doi: 10.30871/jaic.v9i3.9584.

[13] S. Pudjiati, A. N. Fachmi, R. D. Patria, F. R. Ramadhan, and A. P. Sari, "Perbandingan Kinerja Klasifikasi Data Mining untuk Diagnosis Diabetes," *J. Media Inform.*, vol. 7, no. 1 SE-Articles, pp. 384–391, Feb. 2026, doi: 10.55338/jumin.v7i1.8024.

[14] I. Akbar, F. Supriadi, and D. Indra Junaedi, "PEMANFAATAN MACHINE LEARNING DI BIDANG KESEHATAN," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 9, no. 1, pp. 1744–1749, Jan. 2025, doi: 10.36040/jati.v9i1.12663.

[15] A. A. Arifiyanti and E. D. Wahyuni, "SMOTE: METODE PENYEIMBANG KELAS PADA KLASIFIKASI DATA MINING," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, Feb. 2020, doi: 10.33005/scan.v15i1.1850.

[16] A. Nugroho and E. Rilvani, "Penerapan Metode Oversampling SMOTE Pada Algoritma Random Forest Untuk Prediksi Kebangkrutan Perusahaan," *Techno.Com*, vol. 22, no. 1, pp. 207–214, Feb. 2023, doi: 10.33633/tc.v22i1.7527.

[17] N. Nanda Pradana, A. Agung Subekti, and E. Rilvani, "DETEKSI TRANSAKSI MENCURIGAKAN MENGGUNAKAN DECISION TREE DAN LOGISTIC REGRESSION DENGAN MITIGASI KETIDAKSEIMBANGAN KELAS," *J. Media Akad.*, vol. 3, no. 8 SE-Articles, 2025, doi: 10.62281/v3i8.2680.

[18] Z. Rozikin and J. J. Hidayat, "Perbandingan Metode Oversampling SMOTE dan ADASYN pada Klasifikasi Diabetes Menggunakan Algoritma CatBoost," *J. Manaj. Inform. Teknol.*, vol. 6, no. 1, pp. 151–164, 2026, doi: 10.51903/mifortekh.v6i1.1157.

[19] Alwaliyanto, Siska Kurnia Gusti, Iis Afrianty, and Fadhilah Syafria, "Penerapan Metode ADASYN Dalam Mengatasi Imbalanced Data Untuk Klasifikasi Penyakit Stroke Menggunakan Support Vector Machine," *Bull. Comput. Sci. Res.*, vol. 5, no. 4, pp. 532–541, Jun. 2025, doi: 10.47065/bulletincsr.v5i4.612.

[20] N. Surojudin, S. Butsianto, and A. Firmansyah, "Perbandingan Kinerja Naïve Bayes dengan dan Tanpa SMOTE untuk Klasifikasi Gangguan Kecemasan Mahasiswa pada Data Tidak Seimbang," *Bull. Comput. Sci. Res.*, vol. 6, no. 2 SE-, pp. 804–812, Feb. 2026, doi: 10.47065/bulletincsr.v6i2.1021.

[21] A. Nugroho, Wiyanto, and D. Maulana, "COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS IN HANDLING IMBALANCED DATA WITH SMOTE OVERSAMPLING APPROACH," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 11, no. 2, pp. 487–495, Nov. 2025, doi: 10.33480/jitk.v11i2.6956.

[22] X. Wang *et al.*, "Diabetes mellitus early warning and factor analysis using ensemble Bayesian networks with SMOTE-ENN and Boruta," *Sci. Rep.*, vol. 13, no. 1, p. 12718, Aug. 2023, doi: 10.1038/s41598-023-40036-5.

[23] C. Yang, E. A. Fridgeirsson, J. A. Kors, J. M. Reps, and P. R. Rijnbeek, "Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data," *J. Big Data*, vol. 11, no. 1, p. 7, Jan. 2024, doi: 10.1186/s40537-023-00857-7.

[24] M. David and N. Qolbi, "Tinjauan Sistematis Metodologi Penelitian Informatika: Tren, Tantangan, dan Arah Masa Depan," *Karapan Netw. J. J. Comput. Technol. Mob. Ad Hoc Netw.*, vol. 2, no. 2, 2026, [Online]. Available: https://ejournal.omahtabing.com/knj/article/view/480

[25] A. M. Syarif, A. Z. Fanani, H. Himawan, and B. Widjajanto, *METODOLOGI PENELITIAN DALAM INFORMATIKA: KONSEP DAN APLIKASI*. Ritecs, 2025.

[26] M. Maftoun *et al.*, "Malicious URL Detection Using Optimized Hist Gradient Boosting Classifier Based on the Grid Search Method," 2026, pp. 353–364. doi: 10.1007/978-3-032-16469-8_22.

[27] U. Khairani, V. Mutiawani, and H. Ahmadian, "Pengaruh Tahapan Preprocessing Terhadap Model Indobert Dan Indobertweet Untuk Mendeteksi Emosi Pada Komentar Akun Berita Instagram," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 4, pp. 887–894, Aug. 2024, doi: 10.25126/jtiik.1148315.

[28] A. Ichwani, R. I. Kesuma, A. Setiawan, I. E. Wicaksono, and R. Hanifah, "Preventing Data Leakage in Classification via Integrated Machine Learning Pipelines: Preprocessing, Feature Transformation, and Hyperparameter Tuning," *J. Tek. Inform.*, vol. 7, no. 1, pp. 391–410, Feb. 2026, doi: 10.52436/1.jutif.2026.7.1.5490.

[29] A. I. K. Akbar and Y. P. Astuti, "Lung Cancer Classification using the Naïve Bayes Method with SMOTE," *SISTEMASI*, vol. 14, no. 6, p. 2954, Nov. 2025, doi: 10.32520/stmsi.v14i6.5607.

[30] Z. Rozikin, A. T. Zy, and A. Z. Kamalia, "Prediksi Ketebalan Powder Coating Menggunakan Algoritma SVM Dan Naïve Bayes ," *Bull. Inf. Technol.*, vol. 4, no. 2 SE-Articles, Jun. 2023, doi: 10.47065/bit.v4i2.687.

[31] M. Zhu *et al.*, "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data," *IEEE Access*, vol. 6, pp. 4641–4652, 2018, doi: 10.1109/ACCESS.2018.2789428.

[32] A. Gosain and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2017, pp. 79–85. doi: 10.1109/ICACCI.2017.8125820.

[33] I. Darmawan, H. Setiaji, L. Sutriani, A. Supriyadi, R. Rizal, and A. Rahmatulloh, "Ensemble Learning for Malware Classification: A Performance Evaluation Using Random Forest and Hist Gradient Boosting," in *2025 International Conference on Information and Communication Technology (ICoICT)*, 2025, pp. 1–6. doi: 10.1109/ICoICT66265.2025.11193116.

[34] B. Liao, T. Zhou, J. Yu, Y. Liu, T. Zhang, and T. Lu, "Enhancing wildfire spread scale prediction with explainable AI: a hist gradient boosting and SHAP-based approach," *Nat. Hazards*, vol. 122, no. 9, p. 378, May 2026, doi: 10.1007/s11069-026-08145-2.

[35] S. K. N. A. Putri, I. Jumiatin, I. Sulistia, N. A. B. Saputra, and N. Wiranda, "Penerapan Hyperparameter Tuning pada Model Klasifikasi untuk Prediksi Risiko Penyakit Jantung : Implementation of Hyperparameter Tuning for Classification Models in Heart Disease Risk Prediction," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 5, no. 4 SE-, pp. 1181–1189, Oct. 2025, doi: 10.57152/malcom.v5i4.2138.

[36] M. B. Ilmi and Kusrini, "Perbandingan Kinerja Algoritma Machine Learning dalam Deteksi Potensi Risiko HIV," *Buffer Inform.*, vol. 11, no. 1, pp. 36–46, Apr. 2025, doi: 10.25134/buffer.v11i1.355.

[37] J. J. Hidayat, C. Setyowati, and A. P. Werdana, "Perancangan Sistem Prediksi Penyakit pada Tanaman Padi Berbasis Image Processing Menggunakan Algoritma VGG-16 Transfer Learning dan K-Means Segmentation," *J. Pract. Comput. Sci.*, vol. 5, no. 1, pp. 1–15, May 2025, doi: 10.37366/jpcs.v5i1.5759.

[38] A. M. Rifai, S. Raharjo, E. Utami, and D. Ariatmanto, "Analysis for diagnosis of pneumonia symptoms using chest X-ray based on MobileNetV2 models with image enhancement using white balance and contrast limited adaptive histogram equalization (CLAHE)," *Biomed. Signal Process. Control*, vol. 90, p. 105857, Apr. 2024, doi: 10.1016/j.bspc.2023.105857.

[39] A. Amali, D. Maulana, E. Widodo, A. Firmansyah, and M. Danny, "Sentiment Analysis of Bekasi Floods Using SVM and Naive Bayes with Advanced Feature Selection," *Brill. Res. Artif. Intell.*, vol. 4, no. 1, pp. 362–371, Jul. 2024, doi: 10.47709/brilliance.v4i1.4268.

[40] A. Halim and A. Safuwan, "Analisis Sentimen Opini Warganet Twitter Terhadap Tes Screening Genose Pendeteksi Virus Covid-19 Menggunakan Metode Naïve Bayes Berbasis Particle Swarm Optimization," *J. Inform. Teknol. dan Sains*, vol. 5, no. 1, pp. 170–178, 2023.

[41] J. J. Hidayat, A. H. Anshor, and M. S. Anwar, "Pemodelan Deteksi dan Klasifikasi Fraktur Tulang pada Radiografi X-Ray Menggunakan YOLOv8 dan Preprocessing CLAHE," *J. FASILKOM*, vol. 16, no. 1, pp. 31–45, 2026, doi: 10.37859/jf.v16i1.11241.