Journal of Applied Informatics and Computing (JAIC)
Vol.9, No.6, December 2025, pp. 3146~3156
e-ISSN: 2548-6861

Hypertension Risk Prediction Using Stacking Ensemble of CatBoost, XGBoost, and LightGBM: A Machine Learning Approach

Abisakha Saif Alfath [1]*, Ajie Kusuma Wardhana [2]*, Rumini [3]*
* Informatika, Universitas Amikom Yogyakarta
abisakha_saif_alfath@students.amikom.ac.id [1], ajiekusuma@amikom.ac.id [2], rumini@amikom.ac.id [3]

Article Info

Article history:

Received 2025-07-21
Revised 2025-10-09
Accepted 2025-11-05

Keyword:

Hypertension Prediction,
Stacking Ensemble,
Machine Learning,
Imbalanced Data Handling,
Classification Metrics.

ABSTRACT

Hypertension is a leading cause of cardiovascular diseases, chronic kidney failure, and strokes, affecting millions worldwide. Early detection and accurate risk prediction are crucial for effective management and prevention. This study aims to evaluate and compare the performance of different algorithms for predicting hypertension risk using a stacking ensemble approach. The model combines three gradient boosting algorithms XGBoost, LightGBM, and CatBoost as base learners, with Logistic Regression as the meta learner. The dataset, sourced from Kaggle, contains 4,240 instances with demographic and clinical attributes relevant to hypertension. The preprocessing steps included imputing missing values using the median, removing residual null entries, and addressing class imbalance through the SMOTE algorithm. Data were divided into 80% for training and 20% for testing. The evaluation showed that the stacking ensemble model achieved an overall accuracy of 92,65%, with precision, recall, and F1-scores consistently reaching 0.92 for both classes. The confusion matrix revealed minimal misclassification, indicating the model's strong ability to differentiate between low and high risk individuals. These results emphasize that the primary goal of this research is to identify which algorithm provides the best performance for hypertension risk prediction. By evaluating and comparing different models, this study offers insights into choosing the most effective algorithm for clinical decision-making and early detection strategies.

I. PENDAHULUAN

Hipertensi merupakan kondisi medis yang ditandai oleh tekanan darah sistolik dan diastolik yang tinggi di atas tekanan darah normal[1]. Hipertensi bekerja secara perlahan merusak berbagai organ dalam tubuh, seperti ginjal, jantung, mata, dan otak yang menjadi faktor utama penyebab penyakit jantung, gagal ginjal kronis, serta stroke[2].

Merujuk pada data dari World Health Organization (WHO) tahun 2023 diperkirakan 1,28 miliar orang dewasa dalam rentang usia 23-79 tahun di seluruh dunia menderita Hipertensi berpenghasilan rendah dan menengah. Jumlah ini meningkat tajam dari 594 juta pada tahun 1975, dan diperkirakan akan mencapai 1,5 miliar kasus pada tahun 2025[3]. Selain itu, menurut data Kementerian Kesehatan tahun 2023 penderita hipertensi di Indonesia mencapai 34,1% atau sekitar 70 juta penduduk[4]. Hal tersebut mengindikasikan 1 dari 3 orang Indonesia menderita hipertensi.

Berbagai studi sebelumnya mengenai hipertensi menegaskan urgensi penanganan yang cepat dan tepat terhadap kondisi ini. Hipertensi yang tidak terkendali berpotensi menimbulkan komplikasi berat pada organ-organ vital, seperti jantung, ginjal, otak, serta mata[5]. Kondisi menjadi salah satu pemicu utama penyakit jantung koroner dan gagal jantung, yang dalam banyak kasus dapat menyebabkan kematian mendadak apabila tidak ditangani secara menyeluruh[6]. Tekanan darah tinggi yang berlangsung dalam jangka panjang dapat merusak sistem kardiovaskular secara permanen serta meningkatkan kemungkinan terjadinya stroke dan gangguan ginjal kronis[5]. Dengan demikian, hipertensi merupakan penyakit

yang memerlukan perhatian serius, karena keterlambatan dalam proses deteksi dan pengobatan dapat mengarah pada kondisi yang mengancam jiwa seperti gagal jantung, gangguan irama jantung, hingga henti jantung[7].

Sejalan dengan urgensi deteksi dini hipertensi, pemanfaatan Machine Learning muncul sebagai pendekatan inovatif yang berkembang pesat dalam beberapa tahun terakhir. Kemampuan metode Machine Learning dalam menganalisis data medis yang kompleks memungkinkan proses pengambilan keputusan menjadi lebih cepat, akurat, dan berbasis data[2]. Dalam praktiknya, Machine Learning memainkan peran strategis dalam meningkatkan efektivitas skrining serta mendukung upaya pencegahan yang lebih tepat dan terarah[8].

Sejumlah penelitian sebelumnya telah membahas topik yang sejalan dengan studi ini, terutama dalam penggunaan Machine Learning untuk memprediksi risiko hipertensi. Salah satunya membangun model prediktif menggunakan algoritma Artificial Neural Network (ANN) dengan akurasi mencapai 85%, namun tanpa membandingkannya dengan algoritma lain dan tanpa menerapkan metode penyeimbangan data, sehingga berpotensi menimbulkan bias terhadap kelas yang dominan[9]. Studi lain memanfaatkan algoritma XGBoost dan memperoleh hasil akurasi sebesar 88,8%, recall 97,04%, dan F1-score 93,18%, namun cakupan dataset yang terbatas menjadi hambatan dalam menggeneralisasi hasil model secara luas[8]. Sementara itu, penelitian berbasis data skrining Kesehatan menunjukan bahwa algoritma Random Forest mencapai sensitivitas 81,8% dan spesifisitas 62,9%, meskipun masih terdapat ketimpangan performa antar metrik evaluasi[[6]. Di sisi lain, meta-analisis menyimpulkan bahwa secara keseluruhan, model Machine Learning memiliki nilai C-statistic sebesar 0,76, hanya sedikit lebih tinggi dari model regresi konvensional (0,75)[7].

Terdapat beberapa kekurangan dalam penelitian terdahulu, seperti belum optimalnya pengolahan data kompleks, keterbatasan jumlah atribut yang digunakan, serta minimnya implementasi pendekatan ensemble, menunjukkan perlunya strategi prediktif yang lebih menyeluruh. Model klasifikasi tunggal umumnya memiliki keterbatasan dalam mengenali pola data yang bervariasi, serta rentan terhadap permasalahan overfitting, terutama saat digunakan pada data kesehatan yang bersifat heterogen seperti hipertensi[10]. Dalam hal ini, metode stacking ensemble menjadi salah satu alternatif yang menjanjikan karena mampu memadukan performa dari beberapa algoritma dasar untuk menghasilkan prediksi yang lebih konsisten dan presisi. Salah satu keunggulan utama dari stacking adalah kemampuannya dalam mengintegrasikan keluaran berbagai base learners melalui sebuah meta learner, sehingga menghasilkan klasifikasi yang lebih unggul[11].

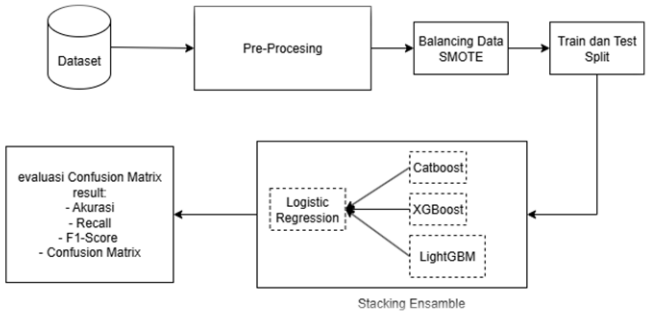
Pemilihan algoritma gradient boosting yang mencakup XGBoost, LightGBM, dan CatBoost didasarkan pada temuan empiris yang menegaskan keunggulannya dalam pengolahan data tabular yang menjadi dasar penelitian ini. Penelitian menunjukkan bahwa Gradient Boosting Decision Trees (GBDT) secara konsisten memiliki kinerja lebih baik dibandingkan model pembelajaran mesin konvensional maupun arsitektur deep learning, sekaligus menawarkan efisiensi komputasi yang lebih tinggi pada berbagai tugas klasifikasi medis berbasis data tabular [12]. Selaras dengan hasil tersebut, studi lain menekankan bahwa varian utama GBDT yakni XGBoost, LightGBM, dan CatBoost masing-masing memiliki keunggulan khusus, seperti: XGBoost unggul dalam akurasi prediksi, LightGBM lebih efisien dari sisi waktu pelatihan, sedangkan CatBoost efektif dalam menangani variabel kategorikal [13]. Oleh karena itu, dasar teoritis dan bukti empiris ini memperkuat argumentasi penggunaan algoritma gradient boosting sebagai komponen inti dalam pendekatan stacking ensemble pada penelitian ini.

Penelitian ini berfokus pada penilaian komparatif untuk menentukan algoritma yang paling unggul di antara model yang digunakan. Dengan demikian, tujuan utama penelitian ini adalah menilai dan mengidentifikasi algoritma dengan performa terbaik untuk prediksi risiko hipertensi.

II. METODE

A. Kerangka Kerja Penelitian

Tahap metode penelitian merupakan tahap yang mencakup alur atau langkah-langkah yang terencana dan logis dari awal hingga akhir untuk memastikan bahwa hasil penelitian dapat dipercaya dan valid. Dengan menggunakan bahasa pemrograman phyton dan software Google Collaboratory. Tahap penelitian memiliki alur atau langkah-langkah pada diagram alir seperti ditunjukkan pada gambar 1.



Gambar 1. Tahapan Penelitian

Penelitian ini difokuskan pada pengembangan model klasifikasi prediktif dengan menerapkan pendekatan stacking ensemble berbasis algoritma pembelajaran mesin modern. Proses dimulai dengan tahap akuisisi dan pra-pemrosesan data, yang melibatkan pembersihan terhadap data yang mengandung nilai kosong, duplikat, dan ketidakkonsistenan yang berpotensi menurunkan akurasi model. Setelah proses tersebut, dataset dibagi menjadi dua bagian, yaitu data pelatihan dan data pengujian, melalui teknik train-test split guna menjamin bahwa evaluasi performa dilakukan secara objektif terhadap data yang belum pernah dilatih sebelumnya.

Langkah ini menjadi elemen krusial untuk menjaga validitas proses pembelajaran mesin yang akan diterapkan.

Tahapan berikutnya melibatkan pembangunan model prediksi dengan teknik stacking ensemble, di mana beberapa algoritma dasar (base learners) seperti CatBoost, XGBoost, dan LightGBM digunakan secara bersamaan untuk menghasilkan output prediksi awal. Hasil prediksi dari model-model tersebut kemudian digabungkan oleh meta-learner Logistic Regression untuk memperoleh hasil akhir yang lebih optimal. Evaluasi performa dilakukan dengan menggunakan sejumlah metrik evaluatif seperti akurasi, recall, dan F1-score, serta dilengkapi dengan interpretasi melalui confusion matrix. Pendekatan evaluasi ini memberikan gambaran menyeluruh terkait kapabilitas model dalam melakukan klasifikasi, serta menegaskan efektivitas teknik ensemble yang digunakan dalam meningkatkan performa prediktif secara keseluruhan.

B. Dataset

Dataset merupakan kumpulan data yang terorganisir dalam format tertentu dan digunakan dalam analis, pelatihan model, penelitian dan sebagai informasi. Dataset Hypertension risk merupakan dataset yang digunakan dalam penelitian ini. Data ini diambil dari dataset Kagle[14]. Dataset ini terdiri atas 4.240 entri dengan 13 fitur, yang mencakup 12 variabel independen berupa atribut demografi dan kesehatan, serta 1 variabel dependen sebagai label kelas untuk memprediksi risiko hipertensi (0 = risiko rendah, 1 = risiko tinggi). Atribut yang tersedia antara lain jenis kelamin, usia, kebiasaan merokok (status perokok dan jumlah rokok per hari), penggunaan obat tekanan darah tinggi (BPMeds), riwayat diabetes, kadar kolesterol total, tekanan darah sistolik dan diastolik, indeks massa tubuh (BMI), detak jantung, serta kadar glukosa. Dengan cakupan informasi tersebut, dataset ini memberikan gambaran komprehensif mengenai faktor-faktor yang berkontribusi terhadap hipertensi dan relevan digunakan dalam pembangunan model prediksi untuk mendukung upaya deteksi dini dan strategi pencegahan. Rincian atribut yang digunakan dalam penelitian ini disajikan pada Tabel I.

TABEL I
FITUR PADA DATASET YANG DIGUNAKAN

Nama Fitur	Tipe Data
Male	Int64
Age	Int64
CurrentSmoker	Int64
CigsPerDay	Float64
BPMeds	Float64
Diabetes	Int64
TotChol	Float64
SysBP	Float64
DiaBP	Float64
BMI	Float64
HeartRate	Float64
Glucose	Float64
Risk	Int64

C. Preprocessing Data

Data mentah yang ada pada bab sebelumnya tidak dapat diproses oleh mesin secara langsung. Perlu dilakukan modifikasi pada data mentah agar menjadi data yang siap diproses oleh mesin[11]. Adapun beberapa tahapan preprocessing.

Pada tahapan awal preprocessing adalah dengan melakukan pengecekan missing value atau data kosong yang ada pada data mentah (Handling Missing Value). Data kosong ini biasanya terjadi karena adanya kesalahan dalam input data atau data tersebut sengaja tidak diisi. Mesin tidak dapat memproses data yang tidak bernilai. Oleh sebab itu perlu dilakukan imputasi missing value atau pengisian data kosong. Terdapat beberapa cara untuk melakukan pengisian data kosong diantaranya adalah menggunakan nilai median. Hal ini dikarenakan data yang kosong melebihi 5% dari keseluruhan data. Imputasi menggunakan nilai median ialah metode efektif untuk menangani data yang hilang, khususnya pada dataset distribusi tidak normal, karena nilai median lebih tahan terhadap pengaruh nilai ekstrim dibandingkan rata-rata(mean). Kemudian teknik preprocessing yaitu menghapus atau mendrop data kosong. Hal ini dilakukan setelah dilakukan teknik imputasi menggunakan median, data yang kosong berjumlah di bawah 5%. Teknik drop digunakan karena jumlah data yang hilang kurang dari 5%, karena tingkat kehilangan data yang rendah ini tidak secara signifikan mempengaruhi estimasi hasil penelitian atau meningkatkan risiko bias [15].

Tahap selanjutnya adalah penanganan ketidakseimbangan kelas pada data (imbalanced data handling) yang bertujuan untuk menyamakan distribusi antar kelas (Balancing Data SMOTE). Dalam penelitian ini digunakan pendekatan oversampling dengan metode Synthetic Minority Oversampling Technique (SMOTE). SMOTE bekerja dengan cara menghasilkan sampel sintetis baru untuk kelas minoritas, bukan sekadar melakukan duplikasi data. Proses ini dilakukan dengan memilih secara acak suatu sampel dari kelas minoritas, kemudian mengidentifikasi sejumlah k tetangga terdekatnya menggunakan algoritma k-nearest neighbors. Sampel sintetis selanjutnya dibentuk dengan melakukan interpolasi linier antara sampel awal dengan salah satu tetangga terdekat tersebut. Dengan cara ini, SMOTE tidak hanya menambah jumlah data pada kelas minoritas, tetapi juga memperluas ruang representasi fitur sehingga distribusi data menjadi lebih seimbang [16]. Teknik ini memiliki kelebihan mengatasi overfitting dan Lebih cocok untuk menangani ketidakseimbangan kelas, terutama jika jumlah sampel kelas minoritas sangat kecil.

D. Train dan Test Split

Train dan Test Split adalah teknik yang umum digunakan dalam evaluasi model machine learning untuk membagi dataset menjadi dua bagian. Bagian pertama Adalah pelatihan (training set) dan bagian kedua untuk pengujian (test set)[17] . Metode ini memungkinkan peneliti untuk melatih model pada satu subset data dan kemudian menguji performa model

tersebut pada subset data yang terpisah, sehingga memberikan gambaran yang lebih akurat tentang bagaimana model akan berfungsi pada data baru yang belum pernah dilihat sebelumnya

E. Algoritma Model

Stacking ensemble merupakan salah satu pendekatan dalam ensemble learning yang menggabungkan hasil prediksi dari berbagai model dasar (base learners) melalui sebuah model tambahan yang dikenal sebagai meta learner, dengan tujuan memperoleh prediksi akhir yang lebih tepat dan andal. Berbeda dengan metode ensemble lain seperti bagging dan boosting yang menggabungkan prediksi secara langsung, stacking mempelajari pola penggabungan terbaik dari output model-model dasar tersebut agar dapat meningkatkan kinerja secara keseluruhan.

Pada implementasinya, stacking memanfaatkan beberapa algoritma base learners yang berbeda, seperti CatBoost, XGBoost, dan LightGBM, yang terkenal mampu menangani data numerik maupun kategorikal secara efektif dan memberikan hasil prediksi yang kuat. Prediksi dari base learners ini selanjutnya dijadikan sebagai input bagi meta learner, biasanya berupa regresi logistik atau model sederhana lain yang bertugas mencari kombinasi optimal dari output tersebut untuk menghasilkan prediksi final [11].Pendekatan ini efektif dalam mengatasi keterbatasan masing-masing base learner, mengurangi risiko overfitting, dan meningkatkan akurasi model, terutama dalam konteks prediksi penyakit kronis seperti hipertensi.

1) *Algoritma Base Learner*

Base learner adalah model dasar yang digunakan dalam teknik ensemble learning, khususnya dalam stacking ensemble, untuk menghasilkan prediksi individual yang kemudian digabungkan oleh meta learner. Pemilihan algoritma base learner yang tepat sangat penting karena berpengaruh langsung terhadap performa akhir model ensemble. Beberapa algoritma berbasis pohon keputusan yang populer dan banyak digunakan sebagai base learners adalah CatBoost, LightGBM, dan XGBoost, karena kemampuan mereka yang unggul dalam mengelola data numerik dan kategorikal serta efisiensi komputasi yang tinggi [18], [19].

CatBoost adalah algoritma boosting berbasis pohon yang dikembangkan untuk secara khusus menangani data kategorikal tanpa perlu konversi eksplisit seperti one-hot encoding. Algoritma ini menggunakan teknik ordered boosting dan pengkodean target statistik yang mencegah overfitting dan meningkatkan akurasi model [20]. CatBoost mampu memberikan hasil prediksi yang akurat dan stabil dengan waktu pelatihan yang efisien, sehingga banyak diaplikasikan dalam berbagai bidang termasuk prediksi risiko penyakit kronis [19].

LightGBM adalah framework gradient boosting yang dioptimalkan untuk kecepatan dan penggunaan memori yang efisien. LightGBM menggunakan metode leaf-wise dengan pembatasan depth yang membantu mengurangi kesalahan prediksi dan overfitting pada dataset besar dan kompleks. Algoritma ini juga mendukung fitur kategorikal secara langsung dan dirancang untuk skala besar, menjadikannya pilihan populer dalam aplikasi machine learning skala industri [20].Penggunaan LightGBM sebagai base learner dalam stacking ensemble telah terbukti meningkatkan performa model prediksi penyakit [18].

XGBoost (Extreme Gradient Boosting) adalah salah satu algoritma boosting yang paling banyak digunakan karena performanya yang tinggi dan kemampuannya untuk menangani dataset besar dengan fitur yang kompleks. XGBoost menggabungkan regularisasi L1 dan L2 untuk mengontrol overfitting serta menggunakan metode tree boosting yang efisien secara komputasi. Algoritma ini telah diaplikasikan secara luas dalam berbagai kompetisi dan studi terkait prediksi kesehatan karena kemampuannya menghasilkan model yang akurat dan robust [19], [20].

2) *Algoritma Meta Learner*

Dalam teknik stacking ensemble, meta learner berperan untuk menggabungkan prediksi dari beberapa base learners guna menghasilkan prediksi akhir yang memiliki tingkat akurasi lebih tinggi. Salah satu algoritma yang paling umum digunakan sebagai meta learner adalah regresi logistik. Algoritma ini merupakan metode statistik yang sering dipakai untuk klasifikasi biner dengan cara memodelkan probabilitas suatu peristiwa berdasarkan variabel input yang diberikan[20].

Regresi logistik bekerja dengan menggunakan fungsi sigmoid untuk menghubungkan probabilitas keluaran dengan kombinasi linier dari variabel input. Model ini menghasilkan output dalam rentang nilai antara 0 dan 1, yang selanjutnya dipakai untuk menentukan kelas prediksi akhir. Secara matematis, probabilitas prediksi kelas positif pada stacking ensemble menggunakan regresi logistik sebagai meta learner.

F. Evaluasi

Evaluasi kinerja model merupakan langkah krusial dalam pengembangan sistem pembelajaran mesin, karena bertujuan menghitung sejauh mana model mampu mengklasifikasikan data dengan tepat. Dalam penelitian ini, digunakan lima metrik utama: precision, recall, F1-score, akurasi, dan confusion matrix, yang umum diaplikasikan pada klasifikasi biner. Setiap metrik memiliki fokus evaluasi yang berbeda sehingga saling melengkapi, dan pemilihannya harus mempertimbangkan konteks permasalahan serta potensi ketidakseimbangan data yang mungkin terjadi[21], [22].

Precision mengukur rasio prediksi positif yang benar dibandingkan dengan semua prediksi yang dikategorikan positif, yaitu TP / (TP + FP). Di sisi lain, recall menilai kemampuan model untuk menemukan seluruh kasus positif aktual, yaitu TP / (TP + FN). F1-score digunakan sebagai ukuran kompromi antara precision dan recall—yang sangat relevan pada situasi data tidak seimbang—dengan menghitung 2 × (precision × recall) / (precision + recall).

Kombinasi ketiga metrik ini memberikan evaluasi lebih menyeluruh terhadap kinerja deteksi model[23].

Akurasi tetap menjadi metrik dasar yang digunakan untuk menilai seberapa besar persentase prediksi yang benar, yaitu (TP + TN) / (TP + TN + FP + FN). Namun, dalam kondisi data tidak seimbang, akurasi dapat menyesatkan.

Confusion matrix digunakan sebagai tampilan visual dua dimensi untuk menggambarkan distribusi prediksi benar dan salah per kelas, memberikan informasi mendalam mengenai jenis kesalahan yang terjadi. Informasi tersebut kemudian bisa dimanfaatkan untuk memperbaiki model melalui optimasi hyperparameter atau teknik pembobotan kelas[21].

III. HASIL DAN PEMBAHASAN

A. Data set

Data yang digunakan merupakan data sekunder yang didapatkan dari platform Kaggle. Karena data yang didapatkan berupa dataset sekunder kekurangan dari dataset yang digunakan memiliki kebolehjadian tidak relevan terhadap keadaan di Indonesia.

Variabel-variabel dalam dataset ini mencakup faktor biologis seperti usia, jenis kelamin (male), kadar kolesterol total (totChol), kadar glukosa darah (glucose), dan indeks massa tubuh (BMI). Selain itu, dataset juga melibatkan informasi gaya hidup seperti status merokok (currentSmoker) dan jumlah rokok yang dihisap per hari (cigsPerDay), serta penggunaan obat tekanan darah (BPMeds) dan status diabetes (diabetes). Detak jantung (heartRate) turut dicatat sebagai penunjang analisis kesehatan kardiovaskular. Variabel target atau label, yaitu Risk, menunjukkan klasifikasi risiko hipertensi dengan nilai 1 untuk risiko tinggi dan 0 untuk risiko rendah.

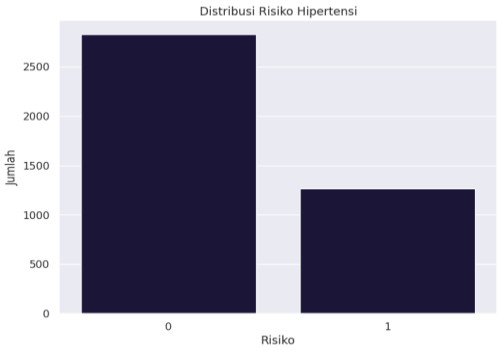
B. Preprocessing

Pada handling missing value ini, terdapat 2 teknik yang diimplementasikan dalam proses ini, pertama menggunakan teknik imputasi median. Teknik ini digunakan untuk fitur glucose yang terdapat nilai kosong sebanyak 388.

Teknik imputasi median dipilih dari pada teknik menghapus data pada kolom glucose, karena data kosong pada fitur ini melebihi 5% dari keseluruhan data. Selain itu, teknik ini memiliki keunggulan untuk menghindari bias, menjaga distribusi data, dan lebih tahan terhadap outlier.

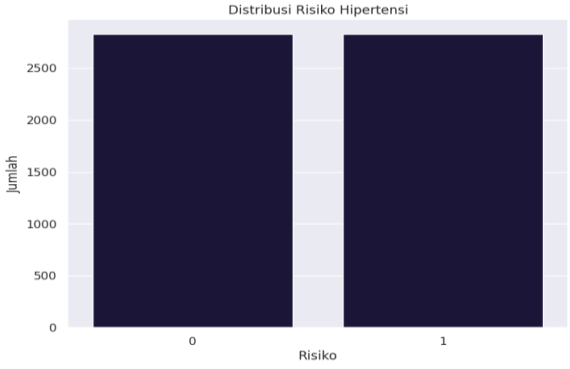
Teknik selanjutnya ialah menghapus data kosong. Teknik ini digunakan pada fitur cigPerDay, BPMeds, totChol dan BMI. Hal ini dilakukan setelah dilakukan teknik imputasi menggunakan median, data yang kosong berjumlah di bawah 5% dari keseluruhan data. Teknik drop digunakan karena jumlah data yang hilang kurang dari 5%, karena tingkat kehilangan data yang rendah ini tidak secara signifikan mempengaruhi estimasi hasil penelitian atau meningkatkan risiko bias .

Tahap berikutnya yaitu melakukan proses handling imbalanced data yang bertujuan untuk Menyeimbangkan Jumlah Kelas dengan menggunakan pendekatan SMOTE. Dapat terlihat pada gambar 4, bahwa distribusi data pada variabel target Risiko dengan nilai 0 untuk risiko rendah dan 1 untuk risiko tinggi tidak seimbang. Jumlah sampel dengan kelas 0 jauh lebih banyak dibandingkan dengan kelas 1. Ketidakseimbangan data ini dapat menyebabkan model cenderung memprediksi kelas mayoritas (kelas 0), sehingga mengurangi akurasi prediksi untuk kelas minoritas (kelas 1), Dapat dilihat pada gambar 4.



Gambar 4. Perbandingan Data.

Untuk mengatasi masalah ini, diperlukan teknik penyeimbangan data, yaitu oversampling pada kelas minoritas menggunakan metode SMOTE (Synthetic Minority Oversampling Technique). SMOTE bekerja dengan cara mensintesis data baru untuk kelas minoritas (kelas 1) berdasarkan interpolasi antara sampel yang ada. Hasilnya, distribusi data menjadi seimbang, dengan jumlah sampel pada kedua kelas (0 dan 1) setara. Dengan distribusi yang seimbang, model memiliki kesempatan yang sama untuk mempelajari pola dari kedua kelas, sehingga mengurangi bias terhadap kelas mayoritas, Dapat dilihat pada gambar 5.



Gambar 5. Hasil Balancing data SMOTE

C. Train dan Test Split

Pada penelitian ini, dataset dengan total 4240 sampel dibagi menjadi dua subset, yaitu data training dan data testing, dengan proporsi 80:20. Pembagian ini dilakukan untuk memastikan bahwa model pembelajaran mesin dapat dilatih dan dievaluasi secara optimal. Data training yang terdiri dari 3272 sampel (80% dari total dataset) digunakan untuk melatih model, mengidentifikasi pola, dan mempelajari hubungan

antara variabel prediktor dan target. Dengan jumlah sampel yang signifikan , data training memberikan representasi yang cukup untuk membangun model yang generalizable.

Data testing, yang mencakup 818 sampel (20% dari total dataset), digunakan untuk mengevaluasi performa model yang telah dilatih. Subset ini tidak pernah dilihat oleh model selama pelatihan, sehingga dapat memberikan estimasi akurasi model secara independen. Proporsi 80:20 ini adalah pembagian standar yang sering digunakan dalam machine learning karena menyediakan cukup data untuk pelatihan sekaligus menjaga representasi dalam evaluasi model. Splitting ini bertujuan untuk memastikan bahwa model dapat memprediksi dengan baik pada data baru yang belum pernah dilihat sebelumnya.

TABEL II
PEMBAGIAN DATA TRAINING DAN DATA TEST

Deskripsi	Data Training	Data Testing
Proporsi	80%	20%
Jumlah	3272	818

D. Algoritma Model

Dengan menggunakan metode Stacking Ensemble untuk klasifikasi, langkah utama yang diambil adalah penyetelan parameter. Hasil penyetelan parameter ditunjukkan pada tabel 3.

TABEL III
HYPERTUNNING PARAMETER

Model	Hyperparameter	Nilai Terbaik
XGBoost	n_estimators	600
	learning_rate	0.05
	max_depth	4
	subsample	0.7
	colsample_bytree	0.7
	gamma	1
	eval_metric	logloss
	use_label_encoder	False
LightGBM	n_estimators	600
	learning_rate	0.01
	max_depth	6
	num_leaves	63
	subsample	1.0
	colsample_bytree	0.7
CatBoost	iterations	600
	learning_rate	0.01
	depth	8
	l2_leaf_reg	1
	verbose	0
Stacking Enemble (Logistic Regression)	Solver	lbfgs
	max_iter	500

Dalam konteks penggunaan algoritma ensemble stacking yang terdiri dari XGBoost, LightGBM, dan CatBoost, proses tuning dilakukan untuk menentukan kombinasi optimal dari hyperparameter pada masing-masing algoritma. Untuk model XGBoost, tuning mencakup parameter seperti jumlah pohon (n_estimators), tingkat pembelajaran (learning_rate), kedalaman maksimum pohon (max_depth), rasio subsample data pelatihan (subsample), fraksi fitur yang digunakan per pohon (colsample_bytree), serta parameter regularisasi gamma. Konfigurasi terbaik yang diperoleh adalah n_estimators sebesar 600, learning_rate sebesar 0.05, max_depth 4, subsample 0.7, colsample_bytree 0.7, dan gamma sebesar 1.

Pada algoritma LightGBM, tuning dilakukan terhadap parameter n_estimators, learning_rate, max_depth, num_leaves, subsample, dan colsample_bytree. Hasil tuning menunjukkan bahwa kombinasi terbaik adalah n_estimators sebesar 600, learning_rate sebesar 0.01, max_depth 6, dan num_leaves 63, dengan nilai subsample dan colsample_bytree masing-masing sebesar 1.0 dan 0.7. Sedangkan pada model CatBoost, parameter yang dituning meliputi jumlah iterasi (iterations), tingkat pembelajaran (learning_rate), kedalaman pohon (depth), dan regularisasi daun (l2_leaf_reg). Konfigurasi terbaik untuk CatBoost diperoleh dengan iterations 600, learning_rate 0.01, depth 8, dan l2_leaf_reg 1.

Ketiga model tersebut kemudian digabungkan ke dalam arsitektur stacking ensemble dengan logistic regression sebagai meta-learner. Logistic regression digunakan untuk menangkap hubungan non-linier dari prediksi yang dihasilkan oleh base learners, dengan konfigurasi solver lbfgs dan max_iter sebanyak 500 iterasi. Proses tuning ini secara keseluruhan bertujuan untuk menyeimbangkan kompleksitas model dengan kemampuannya melakukan generalisasi terhadap data yang belum terlihat. Hasil akhir menunjukkan bahwa konfigurasi tersebut menghasilkan model dengan akurasi yang tinggi dan stabil, serta nilai precision, recall, dan F1-score yang seimbang antar kelas. Dengan demikian, tuning hyperparameter berperan penting dalam meningkatkan performa keseluruhan dari sistem klasifikasi berbasis ensemble.

E. Evaluasi

Penilaian kinerja model adalah langkah penting dalam penelitian ini yang bertujuan untuk menilai sejauh mana model dapat mengklasifikasikan risiko hipertensi dengan tepat. Beberapa metrik utama digunakan dalam proses evaluasi ini untuk memberikan gambaran menyeluruh tentang performa model, seperti akurasi, precision, recall, dan F1-score. Pemilihan metrik-metrik ini didasarkan pada kemampuan mereka untuk secara holistik mengukur keseimbangan antara sensitivitas dan spesifisitas model, terutama dalam situasi data yang tidak seimbang. Hasil dari evaluasi model stacking ensemble dapat dilihat pada tabel 3.

TABEL III
CLASSIFICATION REPORT

Metrix	Nilai
Accuracy	0.9265
Precision	0.93
Recall	0.92
F1-score	0.92

Tabel di atas menyajikan hasil evaluasi kinerja model yang menggunakan beberapa metrik utama, yaitu akurasi, precision, recall, dan F1-score. Secara keseluruhan, model ini menunjukkan performa yang sangat baik dalam

mengklasifikasikan risiko hipertensi, dengan tingkat akurasi sebesar 92,65%. Nilai ini mencerminkan bahwa model berhasil memprediksi dengan tepat lebih dari 92% dari keseluruhan data uji yang diberikan, yang menunjukkan kemampuan model dalam memberikan prediksi yang reliabel.

Metrik precision, yang mengukur proporsi prediksi positif yang benar, tercatat sebesar 0,93. Angka ini mengindikasikan bahwa 93% dari semua instance yang diprediksi sebagai berisiko tinggi benar-benar termasuk dalam kategori tersebut. Hal ini menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam mengidentifikasi individu yang berisiko tinggi hipertensi tanpa menghasilkan banyak prediksi positif yang salah.

Sementara itu, nilai recall yang tercatat sebesar 0,92 menunjukkan kemampuan model dalam mendeteksi 92% dari seluruh instance yang benar-benar berisiko tinggi. Meskipun demikian, terdapat sedikit ketidaktepatan dalam deteksi, yang tercermin dari beberapa instance yang tidak teridentifikasi dengan benar.

Nilai F1-score yang mencapai 0,92 mengindikasikan adanya keseimbangan yang sangat baik antara precision dan recall. F1-score ini memberikan gambaran menyeluruh mengenai performa model, khususnya dalam konteks data yang tidak seimbang. Secara keseluruhan, model stacking ensemble ini terbukti sangat efektif dalam memprediksi risiko hipertensi, dengan kinerja yang konsisten dan stabil di kedua kelas risiko, yakni risiko tinggi dan rendah.

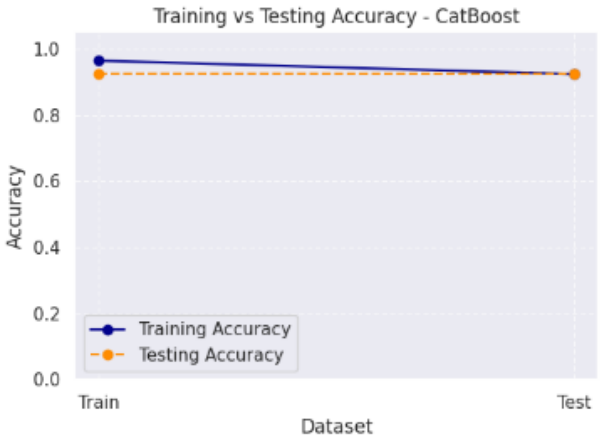
Dibandingkan dengan penelitian-penelitian sebelumnya, menunjukkan bahwa model yang dikembangkan dalam penelitian ini memberikan hasil yang lebih optimal. Salah satu penelitian sebelumnya[24], membangun model prediktif menggunakan algoritma Artificial Neural Network (ANN) dengan akurasi 85%, namun penelitian tersebut tidak membandingkan ANN dengan algoritma lain dan tidak menerapkan metode penyeimbangan data, sehingga berpotensi menimbulkan bias terhadap kelas yang lebih dominan. Studi lainnya[8],menggunakan algoritma XGBoost dan memperoleh akurasi 88,8%, recall 97,04%, dan F1-score 93,18%, namun ukuran dataset yang terbatas menghalangi model ini dalam menggeneralisasi hasil secara lebih luas. Penelitian ini, dengan penerapan teknik stacking ensemble yang menggabungkan XGBoost, LightGBM, dan CatBoost, mencapai akurasi 92,65%, dengan kinerja yang konsisten dan stabil pada kedua kelas risiko hipertensi. Hasil perbandingan di tunjukan pada tabel 4.

TABEL IV
HASIL PERBANDINGAN PENELITIAN

Peneliti	Model	Akurasi
Purwono P et al[24]	ANN	85%
Islam M et al[8]	XGBoost	88,8%
Penelitian ini	Stacking Ensemble(CatBoost, XGBoost, dan LightGBM)	92,65%

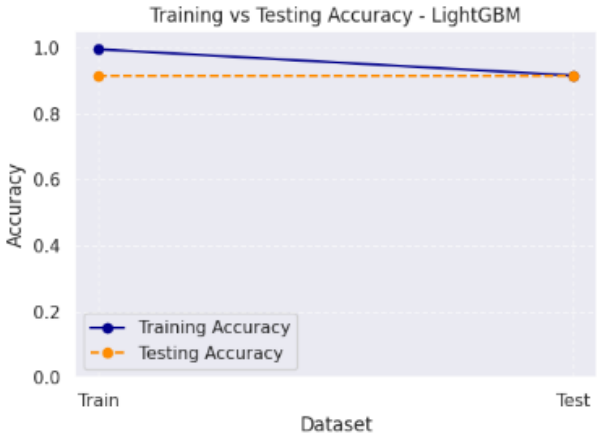
Meskipun penelitian ini menunjukkan hasil yang memuaskan, terdapat beberapa keterbatasan yang perlu diperhatikan. Salah satunya adalah potensi overfitting, terutama karena model stacking ensemble yang digunakan dapat menjadi sangat kompleks dengan banyaknya algoritma yang digabungkan berisiko menyesuaikan terlalu erat dengan data pelatihan dan kurang mampu menggeneralisasi pada data yang belum pernah dilihat sebelumnya. Selain itu, keterbatasan data juga menjadi perhatian, karena dataset yang digunakan berasal dari satu sumber, yaitu Kaggle, yang mungkin tidak sepenuhnya mewakili kondisi demografis atau geografis tertentu, seperti populasi Indonesia. Ukuran sampel yang digunakan, meskipun cukup besar, tetap memiliki batasan dalam hal representasi kondisi global. Oleh karena itu, validasi eksternal pada dataset yang lebih beragam dan dari lokasi yang berbeda akan sangat diperlukan untuk memastikan model ini dapat diandalkan dalam konteks yang lebih luas.

Potensi overfitting yang telah diuraikan sebelumnya perlu dikaji lebih lanjut melalui analisis visual. Untuk itu, disajikan visual perbandingan akurasi pelatihan dan pengujian pada tiap model algoritma guna menilai secara langsung kestabilan kemampuan generalisasi dan mendeteksi indikasi overfitting. Visualisasi ini memberikan konteks awal terhadap kinerja model pada data yang tidak pernah dilihat sebelumnya.



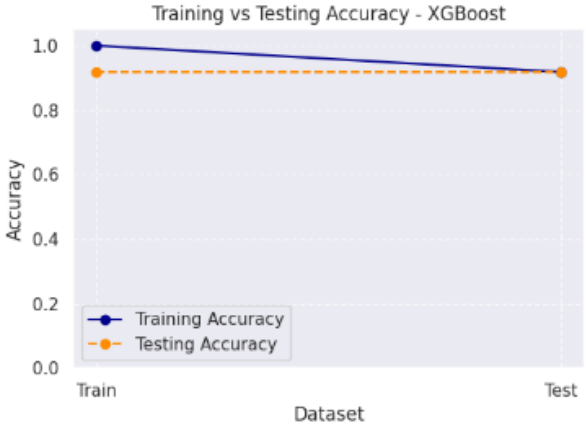
Gambar 6. Training vs Testing Accuracy Catboost

Gambar 6, memperlihatkan perbandingan akurasi pada data latih dan data uji untuk algoritma CatBoost. Akurasi pada data latih hampir mencapai 0,99 sedangkan akurasi pada data uji konsisten sekitar 0,92. Perbedaan yang relatif kecil antara kedua ukuran yang berkisar $\Delta acc \approx 0{,}07–0{,}08$ mengindikasikan rentang generalisasi yang terbatas dan tidak menunjukkan tanda-tanda overfitting yang signifikan. Dengan kata lain, peningkatan kecocokan terhadap data latih tidak memicu penurunan kinerja yang bermakna pada data uji, sehingga menandakan trade off bias varian yang terkelola dengan baik dan mendukung kesimpulan bahwa CatBoost mempertahankan kemampuan generalisasi terhadap sampel yang belum pernah diamati.



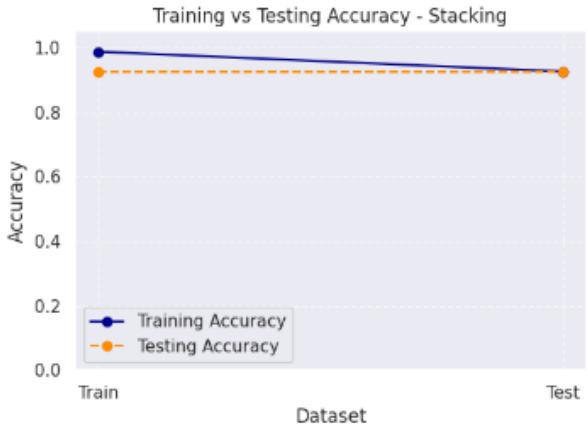
Gambar 7. Training vs Testing Accuracy LightGBM

Gambar 7, menyajikan visualisasi perbandingan akurasi pada data pelatihan dan pengujian untuk algoritma LightGBM. Akurasi pelatihan tercatat sangat tinggi, mendekati 1,00, sementara akurasi pengujian stabil sekitar 0,92. Perbedaan keduanya relatif kecil yang berkisar Δacc ≈ 0,07, yang menunjukkan celah generalisasi sempit dan ketiadaan indikasi overfitting yang bermakna. Temuan ini mengisyaratkan bahwa peningkatan kecocokan model pada data latih tidak disertai degradasi kinerja pada data uji, sehingga trade-off bias–varians berada dalam kendali. Secara keseluruhan, LightGBM memperlihatkan kemampuan generalisasi yang baik terhadap data yang belum pernah diamati.



Gambar 8. Training vs Testing Accuracy XGboost

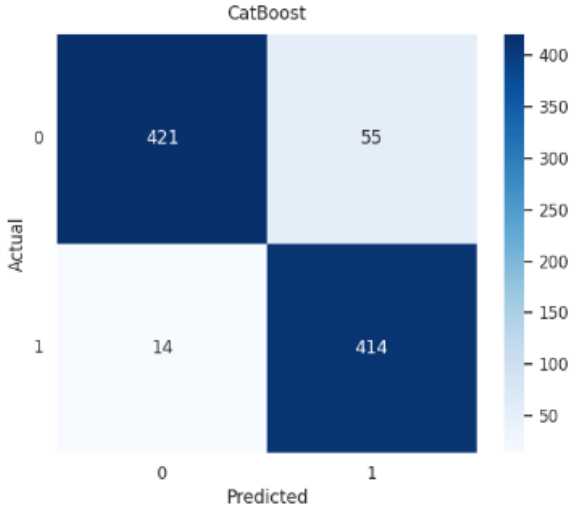
Gambar 8, menunjukkan hasil perbandingan akurasi pelatihan dan pengujian pada algoritma XGBoost. Grafik tersebut memperlihatkan bahwa akurasi pelatihan berada pada tingkat sangat tinggi, mendekati 1,00, sementara akurasi pengujian tetap stabil sekitar 0,92. Perbedaan antara keduanya relatif kecil yang berkisar Δacc ≈0,07–0,08, yang menandakan celah generalisasi sempit dan tidak ditemukan gejala overfitting yang berarti. Hal ini mengindikasikan bahwa meskipun model XGBoost memiliki kecocokan yang kuat terhadap data latih, kinerjanya tidak menurun signifikan ketika diuji pada data baru. Konsistensi ini mempertegas bahwa XGBoost mampu menjaga keseimbangan bias varians dengan baik serta memiliki kapasitas generalisasi yang andal pada data yang belum pernah diamati sebelumnya.



Gambar 9. Training vs Testing Accuracy Stacking Ensemble

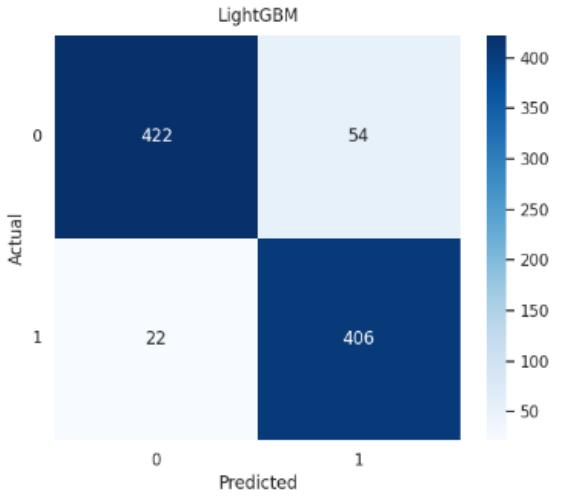
Pada grafik model gambar 9, model stacking yang mengintegrasikan keluaran CatBoost, LightGBM, dan XGBoost menunjukkan akurasi pelatihan sekitar 0,99 dan akurasi pengujian 0,9265. Selisih akurasi yang relatif sempit sebesar Δacc ≈0,06–0,07 mengindikasikan bahwa proses agregasi tidak menambah kompleksitas yang berujung pada overfitting, bahkan mempertahankan dan sedikit meningkatkan kemampuan generalisasi dibandingkan masing-masing model dasar. Kesesuaian nilai akurasi uji pada grafik stacking sebesar 0,9265 dengan hasil evaluasi keseluruhan meneguhkan bahwa model tetap andal pada data yang tidak diamati sebelumnya. Secara keseluruhan, keempat grafik menghadirkan bukti visual dan kuantitatif bahwa selisih akurasi pelatihan dan pengujian yang konsisten. selain itu, selisih antar model pada rentang yang sempit, hal ini menunjukan bahwa model yang dikembangkan dalam penelitian ini tidak mengalami overfitting.

Untuk memberikan gambaran yang lebih jelas mengenai performa model, confusion matrix digunakan sebagai alat bantu visualisasi. Confusion matrix ini menunjukkan distribusi prediksi yang benar dan salah untuk setiap kelas (risiko rendah dan tinggi). Analisis ini memberikan wawasan tentang seberapa baik model dapat membedakan antara kedua kelas tersebut, serta mengidentifikasi kesalahan yang mungkin terjadi pada tiap model.



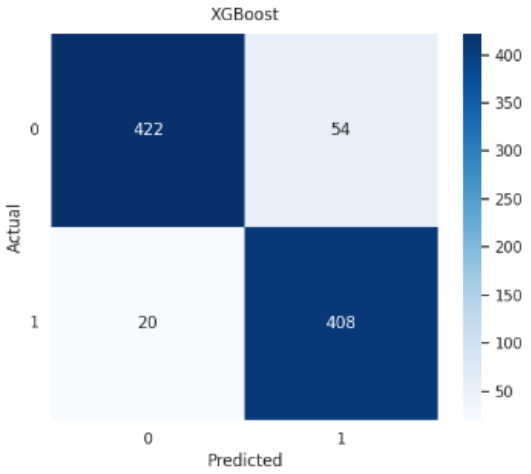
Gambar 10. Confusion Matrix Catboost

Gambar 10, menampilkan hasil confusion matrix untuk CatBoost, menunjukkan distribusi prediksi terdiri atas 421 kasus negatif yang terklasifikasi benar, 55 kasus negatif yang keliru diklasifikasikan sebagai positif, 14 kasus positif yang tidak terdeteksi, dan 414 kasus positif yang terklasifikasi benar. Pola tersebut mencerminkan kepekaan deteksi yang tinggi terhadap kelompok berisiko, terlihat dari jumlah kasus positif terlewat yang paling sedikit serta jumlah kasus positif teridentifikasi yang paling banyak dibandingkan model dasar lainnya. Namun, keunggulan ini disertai peningkatan kesalahan pada kelompok berisiko rendah karena lebih banyak kasus negatif yang salah diberi label positif. Temuan ini menunjukkan bahwa CatBoost menitikberatkan pada penangkapan seluas mungkin terhadap individu berisiko tinggi, dengan konsekuensi adanya tambahan kesalahan pada kelas berisiko rendah. Secara keseluruhan, kinerja yang ditampilkan tetap seimbang.



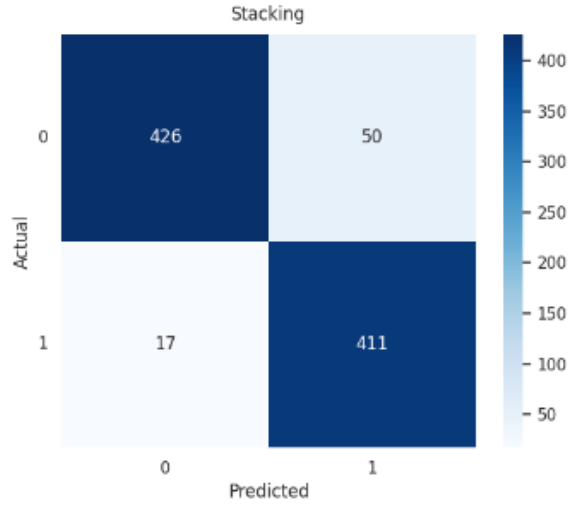
Gambar 11. Confusion Matrix LightGBM

Gambar 11, menyajikan confusion matrix untuk LightGBM, menunjukkan distribusi prediksi berupa 422 kasus negatif yang diklasifikasikan dengan benar, 54 kasus negatif yang keliru diberi label positif, 22 kasus positif yang tidak terdeteksi, serta 406 kasus positif yang teridentifikasi dengan benar. Pola tersebut merefleksikan pendekatan yang lebih berhati-hati dibandingkan CatBoost, terlihat dari meningkatnya jumlah kasus positif yang terlewat sehingga tingkat kepekaan menurun, meskipun kekeliruan terhadap kelompok berisiko rendah sedikit berkurang. Dengan kata lain, LightGBM cenderung bersifat konservatif dalam mengenali individu berisiko tinggi, Sebagian kasus tidak tertangkap namun sekaligus menekan kesalahan pada individu berisiko rendah. Secara menyeluruh, kinerja yang ditampilkan menegaskan kompromi yang berbeda: penurunan kesalahan pada kelompok berisiko rendah dibayar dengan berkurangnya kemampuan penangkapan pada kelompok berisiko tinggi.



Gambar 12. Confusion Matrix XGBoost

Gambar 12, menampilkan confusion matrix untuk XGBoost menunjukkan hasil prediksi dengan 422 kasus negatif yang terklasifikasi benar, 54 kasus negatif yang keliru diidentifikasi sebagai positif, 20 kasus positif yang tidak terdeteksi, serta 408 kasus positif yang terklasifikasi dengan benar. Pola ini memperlihatkan posisi XGBoost di antara CatBoost dan LightGBM, karena jumlah kasus positif yang terlewat lebih sedikit daripada LightGBM tetapi lebih banyak dibandingkan CatBoost, sementara kesalahan pada kasus negatif sebanding dengan LightGBM. Dengan konfigurasi tersebut, XGBoost menampilkan keseimbangan yang relatif baik antara kemampuan mendeteksi individu berisiko tinggi dan menjaga ketepatan pada individu berisiko rendah. Secara keseluruhan, distribusi kesalahan ini menunjukkan bahwa XGBoost mampu mempertahankan performa yang stabil dengan tingkat keseimbangan yang lebih proporsional dibandingkan kedua model dasar lainnya.



Gambar 13. Confusion Matrix Stacking Ensemble

Gambar 13, menampilkan confusion matrix untuk model stacking, menunjukkan distribusi hasil prediksi berupa 426 kasus negatif yang terklasifikasi dengan benar, 50 kasus negatif yang salah dikategorikan sebagai positif, 17 kasus positif yang tidak terdeteksi, serta 411 kasus positif yang terklasifikasi benar. Dengan demikian, pendekatan stacking mampu merestrukturisasi pola kesalahan secara lebih proporsional, menekan kekeliruan pada kelompok berisiko rendah tanpa mengurangi ketepatan dalam mengenali kelompok berisiko tinggi, sehingga menghasilkan kualitas klasifikasi yang lebih unggul dibandingkan setiap model dasar. Hasil ini juga memperlihatkan kinerja yang lebih baik dibandingkan seluruh algoritma penyusunnya, ditandai dengan jumlah kesalahan pada kelompok berisiko rendah yang lebih sedikit serta jumlah kasus negatif yang tepat terklasifikasi paling banyak. Sementara itu, jumlah kasus positif yang terlewat tetap rendah dan jumlah kasus positif yang teridentifikasi dengan benar tetap tinggi. Pola distribusi tersebut menegaskan bahwa stacking tidak hanya memperkuat stabilitas, tetapi juga meningkatkan keseimbangan dalam kinerja klasifikasi.

Secara keseluruhan, meskipun model menunjukkan performa yang cukup memadai, upaya untuk mengurangi kesalahan klasifikasi pada kedua kelas sangat penting untuk meningkatkan akurasi dan efektivitasnya dalam prediksi risiko hipertensi.

IV. KESIMPULAN

Penelitian ini mengembangkan model prediksi risiko hipertensi dengan menggunakan pendekatan stacking ensemble yang menggabungkan algoritma XGBoost, LightGBM, dan CatBoost. Model ini menggunakan dataset sekunder yang diperoleh dari platform Kaggle dan menunjukkan hasil yang sangat baik, dengan akurasi mencapai 92,65%. Selain itu, model ini memiliki kemampuan yang kuat dalam mengklasifikasikan individu berisiko tinggi hipertensi, dengan nilai precision 0,93 dan recall 0,92, serta nilai F1-score yang konsisten sebesar 0,92. Hasil ini membuktikan bahwa model stacking ensemble ini lebih optimal dibandingkan dengan model-model sebelumnya yang menggunakan algoritma seperti Artificial Neural Network (ANN) dan XGBoost, yang masing-masing hanya mencapai akurasi 85% dan 88,8%.

Analisis Training vs Testing Accuracy menunjukkan tidak ada indikasi overfitting pada seluruh model pada peneltian ini, selisih akurasi latih uji konsisten pada rentang Δacc sekitar 0,06–0,08 untuk masing-masing base learner maupun model stacking. Konsistensi ini mengindikasikan kemampuan generalisasi yang baik. Lebih jauh, confusion matrix model stacking (TN 426, FP 50, FN 17, TP 411) menegaskan bahwa FP terendah dan TN tertinggi dicapai tanpa mengorbankan kemampuan menangkap kasus berisiko tinggi (FN tetap rendah), sehingga kualitas klasifikasi lebih kuat dibandingkan tiap base learner secara terpisah.

Untuk memberikan gambaran lebih lanjut mengenai performa model, analisis menggunakan confusion matrix dilakukan, yang menunjukkan bahwa meskipun model ini menunjukkan akurasi yang cukup tinggi, masih ada beberapa kesalahan klasifikasi yang perlu diperbaiki. Model berhasil mengidentifikasi sebagian besar individu dengan risiko tinggi (kelas 1), namun juga menghasilkan kesalahan klasifikasi pada individu dengan risiko rendah (False Positive) dan beberapa individu berisiko tinggi yang tidak teridentifikasi (False Negative). Meskipun demikian, penelitian ini memberikan kontribusi yang signifikan dalam pengembangan model prediksi risiko hipertensi, dan validasi eksternal dengan dataset yang lebih beragam serta upaya pengurangan kesalahan klasifikasi akan meningkatkan keakuratan dan efektivitas model ini lebih lanjut.

DAFTAR PUSTAKA

[1] H. Zhao *et al.*, "Predicting the Risk of Hypertension Based on Several Easy-to-Collect Risk Factors: A Machine Learning Method," *Front Public Health*, vol. 9, Sep. 2021, doi: 10.3389/fpubh.2021.619429.

[2] G. Iaccarino, G. Santulli, H. Tian, Y. Wang, and Y. Zhou, "Development and validation of prediction models for hypertension risks: A cross-sectional study based on n," Sep. 2022. doi: 10.3389/fcvm.2022.928948.

[3] World Health Organization, "Hypertension," Mar. 2023.

[4] Kementerian Kesehatan Republik Indonesia, "Hypertension is called a silent killer, Minister of Health Budi urges routine blood pressure checks," Jun. 2023.

[5] S. M. S. Islam *et al.*, "Machine Learning Approaches for Predicting Hypertension and Its Associated Factors Using Population-Level Data From Three South Asian Countries," *Front Cardiovasc Med*, vol. 9, Mar. 2022, doi: 10.3389/fcvm.2022.839379.

[6] S. Montagna *et al.*, "Machine Learning in Hypertension Detection: A Study on World Hypertension Day Data," *J Med Syst*, vol. 47, no. 1, Dec. 2023, doi: 10.1007/s10916-022-01900-5.

[7] M. Z. I. Chowdhury *et al.*, "Prediction of hypertension using traditional regression and machine learning models: A systematic review and meta-analysis," *PLoS One*, vol. 17, no. 4 April, Apr. 2022, doi: 10.1371/journal.pone.0266334.

[8] M. M. Islam *et al.*, "Predicting the risk of hypertension using machine learning algorithms: A cross sectional study in Ethiopia,"

PLoS One, vol. 18, no. 8 August, Aug. 2023, doi: 10.1371/journal.pone.0289613.

[9] P. Purwono *et al.*, "Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network," *INSECT*, vol. 7, no. 2, p. p, 2022.

[10] S. Reel *et al.*, "Predicting Hypertension Subtypes with Machine Learning Using Targeted Metabolites and Their Ratios," *Metabolites*, vol. 12, no. 8, Aug. 2022, doi: 10.3390/metabo12080755.

[11] A. N. Haya and M. Y. Ramme, "Penerapan Algoritma Stacking Ensemble Machine Learning Berbasis Pohon untuk Prediksi Penyakit Diabetes," *Seminar Nasional Sains Data*, vol. 2024, 2024.

[12] A. Y. Yıldız and A. Kalayci, "Gradient Boosting Decision Trees on Medical Diagnosis over Tabular Data," Aug. 2025, doi: 10.1109/ICAD65464.2025.11114069.

[13] D. Boldini, F. Grisoni, D. Kuhn, L. Friedrich, and S. A. Sieber, "Practical guidelines for the use of gradient boosting for molecular property prediction," *J Cheminform*, vol. 15, no. 1, Dec. 2023, doi: 10.1186/s13321-023-00743-7.

[14] Raihan Khan, "Hypertension-risk-model-main," Kagle.

[15] M. S. Tackney, D. Stahl, E. Williamson, and J. Carpenter, "Missing Step Count Data? Step Away From the Expectation–Maximization Algorithm," *J Meas Phys Behav*, vol. 5, no. 4, pp. 205–214, Dec. 2022, doi: 10.1123/jmpb.2022-0002.

[16] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLoS One*, vol. 12, no. 6, Jun. 2017, doi: 10.1371/journal.pone.0177678.

[17] I. K. Sifat and M. K. Kibria, "Optimizing hypertension prediction using ensemble learning approaches," *PLoS One*, vol. 19, no. 12, Dec. 2024, doi: 10.1371/journal.pone.0315865.

[18] A. N. Haya and M. Y. Ramme, "Penerapan Algoritma Stacking Ensemble Machine Learning Berbasis Pohon untuk Prediksi Penyakit Diabetes," *Seminar Nasional Sains Data*, vol. 2024.

[19] P. Mahajan, S. Uddin, F. Hajati, and M. A. Moni, "Ensemble Learning for Disease Prediction: A Review," Jun. 01, 2023, *MDPI*. doi: 10.3390/healthcare11121808.

[20] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," 2022, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2022.3207287.

[21] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies (Basel)*, vol. 12, no. 7, Jul. 2024, doi: 10.3390/technologies12070113.

[22] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[23] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," *African Journal of Biomedical Research*, pp. 4023–4031, Nov. 2024, doi: 10.53555/ajbr.v27i4s.4345.

[24] P. Purwono *et al.*, "Model Prediksi Otomatis Jenis Penyakit Hipertensi dengan Pemanfaatan Algoritma Machine Learning Artificial Neural Network," vol. 7, no. 2, p. p, 2022.

.