

Stacking Ensemble Learning untuk Klasifikasi Status Gizi Balita Berdasarkan Data Antropometri: Studi Komparatif dan Validasi Lintas *Dataset*

Stacking Ensemble Learning for Toddler Nutritional Status Classification Based on Anthropometric Data: A Comparative Study and Cross-Dataset Validation

Lili Tanti^{a,1}, Budi Triandi^{b,2}, Safrizal^{c,3}

^{a,b}Universitas Potensi Utama, Medan, Indonesia

^cUniversitas Muhammadiyah Asahan, Kisaran, Indonesia

lilitanti82@gmail.com¹, buditriandi@gmail.com², rizalsyl75@gmail.com³

ABSTRAK

Stunting merupakan kondisi kegagalan pertumbuhan linear pada balita akibat akumulasi kekurangan gizi kronis, yang di Indonesia masih menjadi masalah kesehatan masyarakat dengan prevalensi nasional sebesar 19,8% pada tahun 2024. Penelitian ini bertujuan membangun model stacking ensemble untuk klasifikasi status gizi balita berdasarkan data antropometri (usia, jenis kelamin, dan tinggi badan), serta mengevaluasi kemampuan generalisasi model melalui validasi lintas dataset. Metode yang digunakan adalah stacking ensemble dengan Random Forest, Support Vector Machine, dan XGBoost sebagai base learner, serta Logistic Regression sebagai meta-learner. Dua dataset publik digunakan, yaitu dataset A dari Kaggle (120.999 balita) dan dataset B dari Mendeley Data yang bersumber dari Kabupaten Jeneponto, Sulawesi Selatan (39.853 balita setelah pembersihan). Hasil pengujian dalam-dataset pada dataset A menunjukkan model stacking ensemble memperoleh akurasi 99,94% dan F1-score makro 99,90%, sedikit mengungguli Random Forest (99,93%), XGBoost (99,12%), dan SVM (98,37%). Namun, hasil validasi lintas dataset menunjukkan penurunan performa yang signifikan, dengan akurasi 68,80% saat model dilatih pada dataset A dan diuji pada dataset B, serta 79,12% pada arah sebaliknya. Temuan ini mengindikasikan bahwa performa tinggi pada pengujian dalam-dataset belum tentu mencerminkan kemampuan generalisasi model terhadap populasi atau sumber data yang berbeda. Skema validasi lintas dataset yang diterapkan dalam penelitian ini mengungkap adanya kesenjangan generalisasi (*generalization gap*) yang belum banyak dibahas pada studi klasifikasi status gizi balita sebelumnya, sehingga temuan ini dapat menjadi dasar bagi pengembangan model prediksi stunting yang lebih andal di masa mendatang.

Kata Kunci : stunting; klasifikasi status gizi; stacking ensemble; machine learning; validasi lintas dataset

ABSTRACT

Stunting is a linear growth failure condition in toddlers caused by the accumulation of chronic undernutrition, which remains a public health problem in Indonesia with a national prevalence of 19.8% in 2024. This study aims to build a stacking ensemble model for classifying toddler nutritional status based on anthropometric data (age, sex, and height), and to evaluate the model's generalization ability through cross-dataset validation. The method used is stacking ensemble with Random Forest, Support Vector Machine, and XGBoost as base learners, and Logistic Regression as the meta-learner. Two public datasets were used: dataset A from Kaggle (120,999 toddlers) and dataset B from Mendeley Data sourced from Jeneponto Regency, South Sulawesi (39,853 toddlers after cleaning). Within-dataset testing on dataset A showed that the stacking ensemble achieved 99.94% accuracy and a 99.90% macro F1-score, slightly outperforming Random Forest (99.93%), XGBoost (99.12%), and SVM (98.37%). However, cross-dataset validation revealed a substantial performance drop, with 68.80% accuracy when the model was trained on dataset A and tested on dataset B, and 79.12% in the reverse direction. These findings indicate that high within-dataset performance does not necessarily reflect a model's generalization ability across different populations or data sources. The cross-dataset validation scheme applied in this study reveals a generalization gap that has been largely overlooked in previous toddler nutritional status classification studies, providing a foundation for developing more reliable stunting prediction models in the future.

Keywords : stunting; nutritional status classification; stacking ensemble; machine learning; cross-dataset validation

Disubmit: *Info Artikel :* *Direview:* *Diterima :*

Copyright © 2026 - Journal UPU. All rights reserved.

1. PENDAHULUAN

Stunting adalah kondisi kegagalan pertumbuhan linear pada anak akibat akumulasi kekurangan gizi kronis dan infeksi berulang, yang ditandai dengan nilai z-score tinggi badan menurut usia (height-for-age z-score) lebih rendah dari -2 standar deviasi berdasarkan standar pertumbuhan Organisasi Kesehatan Dunia (WHO). Dampak stunting tidak hanya bersifat fisik, tetapi juga berkaitan dengan gangguan perkembangan kognitif dan penurunan produktivitas ekonomi di masa dewasa, sehingga penanganannya menjadi prioritas pembangunan kesehatan global [1].

Berbagai penelitian menunjukkan bahwa faktor penentu stunting bersifat multidimensi, mencakup aspek antropometri, sosioekonomi, dan lingkungan, sehingga diperlukan pendekatan berbasis data untuk mendukung deteksi dini dan intervensi yang lebih tepat sasaran. Komarulzaman dkk. [2] menekankan bahwa pencapaian target penurunan stunting memerlukan pendekatan lintas sektor yang saling terkait pada tingkat kabupaten/kota, mengingat determinan stunting tidak hanya bersumber dari faktor asupan gizi langsung, tetapi juga faktor tidak langsung seperti akses sanitasi, pendidikan ibu, dan kondisi sosioekonomi rumah tangga. Di Indonesia, Novalina dkk. [3] mencatat bahwa meskipun tren prevalensi stunting nasional terus menurun, capaian tersebut masih berada di atas ambang batas yang ditetapkan WHO dan bervariasi cukup besar antarwilayah, sehingga sistem klasifikasi dan prediksi berbasis machine learning yang akurat dan andal menjadi salah satu instrumen pendukung pengambilan keputusan yang penting bagi tenaga kesehatan di tingkat layanan primer.

Perkembangan teknologi machine learning telah banyak dimanfaatkan untuk klasifikasi dan prediksi status gizi balita. Beberapa penelitian menerapkan algoritma tunggal maupun kombinasi algoritma untuk mendeteksi risiko malnutrisi pada anak di berbagai negara. Khan dan Yunus [4] mengembangkan pendekatan ensemble hibrida untuk memprediksi malnutrisi pada balita di negara-negara Afrika Sub-Sahara, sementara Anku dan Duah [5] membandingkan tujuh algoritma machine learning untuk memprediksi undernutrisi pada anak di Ghana. Penelitian serupa oleh Bitew, Sparks, dan Nyarko [6] di Ethiopia menunjukkan bahwa algoritma berbasis pohon memberikan kemampuan prediksi yang lebih baik dibandingkan model generalized linear mixed. Zemariam dkk. [7] menerapkan machine learning dengan teknik SMOTE untuk memprediksi stunting beserta determinan sosioekonominya pada remaja putri di Ethiopia, sedangkan Rao dkk. [8] melakukan meta-analisis terhadap berbagai studi machine learning berbasis data Demographic and Health Surveys untuk memprediksi malnutrisi anak.

Di Indonesia, sejumlah penelitian juga telah membandingkan performa algoritma machine learning untuk klasifikasi status stunting pada balita. Puspita dkk. [9] membandingkan XGBoost dan Random Forest untuk klasifikasi stunting berbasis faktor ekonomi, sementara Elim dan Utami [10] membandingkan SVM dan Random Forest dengan optimasi grid search. Fadil dkk. [11] menerapkan SVM dengan SMOTE untuk prediksi stunting multikelas di Kabupaten Sumedang, sedangkan Hasdina dkk. [12] mengembangkan pendekatan hybrid machine learning untuk klasifikasi, prediksi, dan optimasi clustering stunting di Aceh. Di ranah internasional, Hendy dkk. [13] dan Azriani dkk. [14] membandingkan beberapa algoritma supervised machine learning untuk klasifikasi status gizi pada populasi Mesir dan Indonesia, sementara Siramaneerat dkk. [15] melakukan analisis multilevel terhadap determinan stunting di wilayah urban dan rural Indonesia. Ndagijimana dkk. juga aktif meneliti topik ini di Rwanda, baik melalui pendekatan machine learning konvensional [16] maupun jaringan saraf tiruan

[17], sedangkan Ndagijimana, Elfving, Umubyeyi, dan Lind [18] mengidentifikasi faktor risiko stunting yang dapat dimodifikasi pada berbagai tingkatan populasi di Rwanda Utara.

Penelitian di kawasan lain turut memperkaya bukti empiris terkait prediksi stunting berbasis data. Sur dkk. [19] menerapkan machine learning berbasis data survei kesehatan nasional dan citra satelit untuk memprediksi stunting tingkat rumah tangga di India, sementara Fogarty dkk. [20] menggunakan model XGBoost untuk mengidentifikasi determinan sosial-ekonomi stunting di Meksiko. Begashaw dkk. [21] mengembangkan pendekatan deep learning berbasis LSTM-FC untuk klasifikasi status gizi anak di Ethiopia, dan Chibuye dkk. [22] melakukan tinjauan sistematis mengenai keterkaitan komposisi mikrobioma usus dengan kejadian stunting pada balita. Akuze dkk. [23] turut menyoroti peran data science dan kecerdasan buatan secara lebih luas dalam mendukung kesehatan ibu, bayi baru lahir, dan anak melalui tinjauan cakupan (scoping review).

Selain pendekatan machine learning konvensional, penelitian yang mengintegrasikan model ensemble dan deep learning dengan strategi penyeimbangan kelas juga mulai berkembang, seperti yang dilakukan oleh Richard dkk. [24] untuk deteksi dini stunting dan wasting, serta Lailan dkk. [25] yang mengintegrasikan intervensi berbasis edukasi dengan machine learning untuk pencegahan stunting di Lombok Timur. Pendekatan berbasis deep learning turut dieksplorasi oleh Islam dkk. [26] untuk prediksi malnutrisi pada perempuan di Bangladesh, serta Ankalaki dkk. [27] dan Biradar dan Naik [28] yang menerapkan arsitektur deep learning untuk deteksi malnutrisi pada anak.

Meskipun berbagai penelitian di atas telah menunjukkan performa yang menjanjikan, sebagian besar studi tersebut memiliki dua keterbatasan utama. Pertama, sebagian besar penelitian hanya menggunakan satu sumber dataset dalam skema evaluasi latih-uji tunggal, sehingga belum diketahui secara jelas bagaimana kemampuan generalisasi model tersebut terhadap data dari populasi atau sumber yang berbeda. Kedua, penerapan metode stacking ensemble, yang secara teoretis menggabungkan kekuatan prediktif dari beberapa base learner melalui sebuah meta-learner, masih relatif jarang diterapkan secara khusus pada kasus klasifikasi status gizi balita berbasis data antropometri, meskipun pendekatan ini telah terbukti efektif pada domain kesehatan lain. Ramón dkk. [29] menunjukkan efektivitas XGBoost untuk klasifikasi pasien pada domain kesehatan lain, sedangkan Ghasemieh dkk. [30] membuktikan bahwa model stacking yang menggabungkan beberapa base learner mampu meningkatkan akurasi prediksi readmisi pasien penyakit jantung dibandingkan model tunggal.

Tinjauan sistematis oleh Mahajan dkk. [31] terhadap tren ensemble learning pada prediksi penyakit menunjukkan bahwa model stacking dan hibrida secara konsisten memberikan performa tertinggi dibandingkan pendekatan bagging atau voting sederhana. Temuan serupa dilaporkan oleh Sifat, Kibria, dan Popovic [32] pada prediksi hipertensi, serta oleh Al-Nafjan dkk. [33] dalam tinjauan sistematis mengenai kecerdasan buatan pada layanan kesehatan prediktif, yang menyoroti dominasi model ensemble berbasis pohon seperti Random Forest dan XGBoost untuk data klinis terstruktur. Meskipun demikian, interpretabilitas model masih menjadi tantangan; Ponce-Bobadilla dkk. [34] menyediakan panduan praktis penerapan SHAP (Shapley Additive Explanations) untuk menjelaskan prediksi model machine learning, sedangkan Im, Kim, dan Kim [35] menunjukkan penerapan kombinasi XGBoost dan SHAP untuk meningkatkan interpretabilitas model prediksi pada domain kesehatan masyarakat.

Secara lebih sistematis, sintesis kritis terhadap literatur di atas menunjukkan tiga pola lintas studi. Ditinjau dari cakupan sumber data, mayoritas penelitian terdahulu [4]–[9], [11]–[23], [25]–[28] mengandalkan satu sumber dataset tunggal per studi, baik berupa survei kesehatan nasional, data lapangan berskala kecil, maupun dataset publik hasil kurasi, sehingga kemampuan model dalam menggeneralisasi ke sumber data atau populasi lain belum pernah diuji secara eksplisit. Ditinjau dari pilihan metode, sebagian besar studi terbatas pada penerapan algoritma tunggal atau perbandingan performa antaralgoritma pada level yang setara [9]–[15], sedangkan pendekatan stacking ensemble justru lebih banyak dieksplorasi pada domain kesehatan di luar konteks status gizi balita [29], [30], [32]; penerapannya secara khusus pada data antropometri balita, sebagaimana disoroti oleh tinjauan sistematis

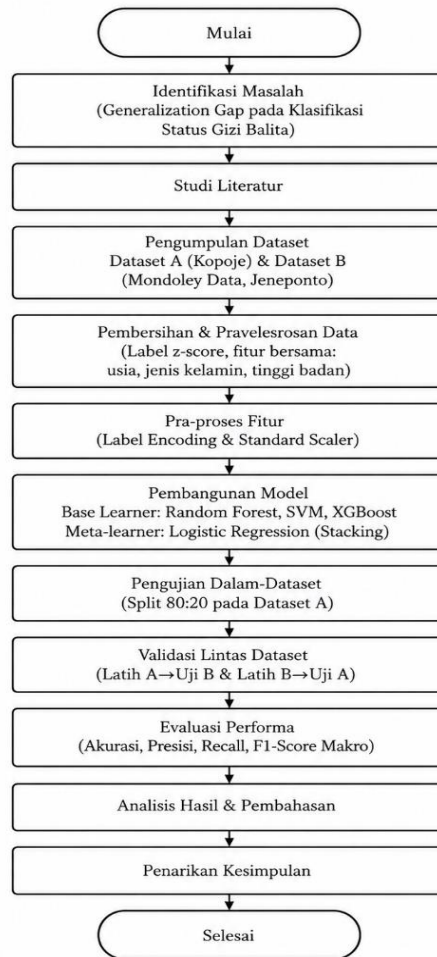
Mahajan dkk. [31], masih jarang dilaporkan. Ditinjau dari skema evaluasi, hampir seluruh studi yang ditelaah menggunakan skema latih-uji tunggal (single train-test split atau k-fold cross-validation) pada satu dataset yang sama, tanpa menyertakan skema pengujian pada dataset independen dari sumber dan populasi yang berbeda. Ketiga pola tersebut, yaitu ketergantungan pada sumber data tunggal, keterbatasan eksplorasi stacking ensemble pada data antropometri balita, dan absennya skema validasi lintas dataset, saling berkaitan dan secara bersama-sama menjadi dasar rumusan masalah serta posisi kebaruan penelitian ini di antara literatur yang ada.

Berdasarkan kesenjangan tersebut, penelitian ini bertujuan untuk: (1) membangun model stacking ensemble yang menggabungkan Random Forest, Support Vector Machine, dan XGBoost sebagai base learner dengan Logistic Regression sebagai meta-learner untuk klasifikasi status gizi balita berdasarkan data antropometri; (2) membandingkan performa model stacking ensemble dengan model tunggal sebagai baseline; dan (3) mengevaluasi kemampuan generalisasi model melalui skema validasi lintas dataset, yaitu dengan melatih model pada satu dataset dan mengujinya pada dataset lain yang berasal dari sumber dan wilayah berbeda. Kontribusi utama penelitian ini adalah menyediakan bukti empiris mengenai kesenjangan generalisasi (generalization gap) pada model klasifikasi status gizi balita, yang belum banyak dibahas pada penelitian-penelitian sebelumnya.

Kebaharuan penelitian ini terletak pada aspek metodologi penelitian serta analisis dan interpretasi data. Dari sisi metodologi, penelitian ini menerapkan skema validasi lintas dataset pada klasifikasi status gizi balita, yaitu melatih model pada satu dataset dan mengujinya pada dataset lain yang berbeda sumber dan wilayah pengumpulan datanya, sebuah pendekatan yang belum diterapkan pada studi-studi sejenis sebelumnya yang umumnya hanya mengevaluasi model pada satu sumber data tunggal. Dari sisi analisis dan interpretasi data, penelitian ini mengungkap adanya kesenjangan generalisasi (generalization gap) yang signifikan antara performa model pada pengujian dalam-dataset dan performa lintas dataset, sebuah temuan dan sudut pandang kritis yang belum banyak diangkat secara eksplisit pada literatur klasifikasi status gizi balita berbasis machine learning.

2. METODE

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tahapan meliputi pengumpulan data, pembersihan dan penyelarasan data, praproses, pembangunan model stacking ensemble, serta evaluasi performa melalui skema pengujian dalam-dataset dan lintas dataset. Tahapan penelitian secara keseluruhan digambarkan pada flowchart berikut.



Gambar 1. Tahapan Penelitian

Gambar 1 menunjukkan alur tahapan penelitian secara berurutan, mulai dari identifikasi masalah terkait kesenjangan generalisasi (generalization gap) pada klasifikasi status gizi balita, studi literatur, pengumpulan dua dataset publik (Dataset A dari Kaggle dan Dataset B dari Mendeley Data), pembersihan dan penyelarasan data, praproses fitur, pembangunan model stacking ensemble (Random Forest, SVM, dan XGBoost sebagai base learner dengan Logistic Regression sebagai meta-learner), pengujian dalam-dataset, validasi lintas dataset, evaluasi performa menggunakan metrik akurasi dan F1-score makro, hingga analisis hasil dan penarikan kesimpulan. Rincian setiap tahapan dijelaskan pada subbagian berikut.

A. Dataset

Penelitian ini menggunakan dua dataset publik. Dataset A bersumber dari platform Kaggle berjudul “Stunting Toddler (Balita) Detection”, terdiri atas 120.999 data balita dengan atribut usia (bulan), jenis kelamin, tinggi badan (cm), dan status gizi yang terbagi ke dalam empat kelas, yaitu severely stunted, stunted, normal, dan tinggi. Dataset B bersumber dari Mendeley Data berjudul “Stunting and Nutritional Status of Toddler, Jeneponto, Sulawesi Selatan”, yang merupakan data lapangan hasil pengukuran antropometri balita di Kabupaten Jeneponto, Sulawesi Selatan, periode 2021–2024, dengan jumlah awal 40.071 data.

Dataset B tidak memiliki label status gizi berkategori empat kelas secara langsung, melainkan berupa nilai z-score tinggi badan menurut usia (Z-Score H/A) yang dihitung menggunakan Persamaan (1) berikut.

$$Z - Score H/A = \frac{X - Median}{SD Baku Rujukan} \quad (1)$$

Dengan X adalah nilai tinggi badan hasil pengukuran, Median adalah nilai median tinggi badan menurut usia dan jenis kelamin pada populasi rujukan WHO, dan SD Baku Rujukan adalah simpangan baku populasi rujukan pada kelompok usia yang sama. Label status gizi pada dataset B kemudian diturunkan dari nilai Z-Score H/A tersebut menggunakan ambang baku antropometri WHO [1], yaitu: z-score kurang dari -3 dikategorikan severely stunted; -3 sampai kurang dari -2 dikategorikan stunted; -2 sampai 3 dikategorikan normal; dan lebih dari 3 dikategorikan tinggi. Sebanyak 216 data dengan nilai z-score kosong dan 2 data dengan usia kosong dibuang dari analisis, sehingga dataset B akhir yang digunakan berjumlah 39.853 data.

Proses pembersihan data juga mengidentifikasi anomali pada kolom berat badan (Weight) dataset B, yaitu sebagian nilai desimal (≈ 20.158 data) yang secara tidak sengaja terkonversi menjadi format tanggal akibat kesalahan interpretasi format sel pada perangkat lunak spreadsheet. Nilai tersebut dipulihkan dengan mengekstraksi kembali komponen hari dan bulan dari nilai tanggal yang tersimpan. Karena kolom berat badan tidak tersedia pada dataset A, kolom ini tidak digunakan dalam skema validasi lintas dataset, namun tetap disertakan pada tahap eksplorasi data.

B. Penyelesaian dan Praproses Fitur

Untuk memungkinkan validasi lintas dataset, hanya fitur yang tersedia pada kedua dataset yang digunakan sebagai fitur bersama (common features), yaitu usia (bulan), jenis kelamin, dan tinggi badan (cm). Fitur kategorikal jenis kelamin dan label target dikodekan menggunakan Label Encoding, sedangkan fitur numerik distandardisasi menggunakan Standard Scaler agar memiliki rata-rata nol dan simpangan baku satu, yang penting terutama bagi algoritma SVM dan Logistic Regression yang sensitif terhadap skala fitur. Pada skema validasi lintas dataset, encoder dan scaler yang telah dilatih (fit) pada dataset sumber diterapkan langsung pada dataset target tanpa dilatih ulang, guna mensimulasikan kondisi penerapan model pada data baru yang belum pernah dilihat sebelumnya.

C. Skema Eksperimen

Penelitian ini menerapkan dua skema evaluasi. Skema pertama adalah pengujian dalam-dataset (within-dataset), yaitu dataset A dibagi menjadi data latih dan data uji dengan proporsi 80:20 secara stratifikasi berdasarkan kelas target menghasilkan 96.799 data latih dan 24.200 data uji. Skema kedua adalah validasi lintas dataset (cross-dataset validation), yaitu model dilatih secara penuh pada satu dataset dan diuji langsung pada dataset lain tanpa pelatihan ulang, dengan dua arah pengujian: (1) model dilatih pada dataset A dan diuji pada dataset B; dan (2) model dilatih pada dataset B dan diuji pada dataset A. Skema kedua ini bertujuan mengukur kemampuan generalisasi model terhadap populasi dan sumber data yang berbeda karakteristik pengumpulannya.

D. Model Stacking Ensemble

Model stacking ensemble yang dibangun terdiri atas tiga base learner, yaitu Random Forest, Support Vector Machine dengan kernel Radial Basis Function, dan XGBoost, serta satu meta-learner berupa Logistic Regression yang mempelajari kombinasi optimal dari probabilitas prediksi ketiga base learner tersebut. Skema stacking ini mengikuti konsep stacked generalization, sejalan dengan tinjauan Mahajan dkk. [31] yang menunjukkan bahwa pendekatan stacking dan hibrida secara konsisten memberikan performa tertinggi pada berbagai kasus prediksi kesehatan, dengan skema cross-validation internal (5-fold) untuk menghindari kebocoran data (data leakage) saat melatih meta-learner. Sebagai perbandingan, ketiga base learner juga dievaluasi secara individual sebagai model baseline. Seluruh model diimplementasikan menggunakan pustaka scikit-learn.

Pemilihan skema stacking dibandingkan pendekatan ensemble lain, yaitu bagging, boosting, dan voting, didasarkan pada beberapa pertimbangan akademik. Bagging (sebagaimana diterapkan pada Random Forest) dan boosting (sebagaimana diterapkan pada XGBoost) pada dasarnya membangun kumpulan base learner yang homogen dari satu algoritma dasar yang sama, masing-masing bertujuan menurunkan varians melalui agregasi bootstrap atau menurunkan bias melalui koreksi kesalahan residual secara sekuensial, sehingga keduanya kurang sesuai apabila tujuan penelitian adalah mengombinasikan algoritma yang berbeda karakteristik. Stacking, sebaliknya, secara khusus dirancang untuk menggabungkan base learner yang bersifat heterogen, dalam penelitian ini yaitu Random Forest (berbasis pohon dan bagging), SVM (berbasis margin dan kernel), dan XGBoost (berbasis pohon dan boosting), yang masing-masing memiliki bias induktif dan asumsi struktural yang berbeda terhadap data, sehingga berpotensi menangkap pola yang saling melengkapi pada ruang fitur antropometri. Dibandingkan dengan voting (baik hard voting maupun soft voting) yang menggabungkan prediksi base learner dengan bobot tetap atau seragam tanpa proses pembelajaran tambahan, stacking menggunakan meta-learner (Logistic Regression) yang mempelajari secara adaptif bobot kontribusi optimal setiap base learner berdasarkan data, sehingga secara teoretis mampu memberikan performa yang lebih baik atau setidaknya setara dibandingkan base learner individualnya maupun skema voting statis. Pertimbangan ini konsisten dengan tinjauan sistematis Mahajan dkk. [31], yang menunjukkan bahwa pendekatan stacking dan hibrida secara konsisten mengungguli bagging maupun voting sederhana pada berbagai kasus prediksi kesehatan, sehingga stacking dipandang sebagai pilihan metodologis yang paling sesuai untuk mengombinasikan kekuatan prediktif Random Forest, SVM, dan XGBoost pada penelitian ini.

Random Forest bekerja dengan membangun sejumlah besar pohon keputusan secara independen melalui teknik bootstrap aggregating (bagging), di mana setiap pohon dilatih pada subset data hasil pengambilan sampel acak dengan pengembalian, serta pemilihan subset fitur secara acak pada setiap pemisahan simpul (split). Prediksi akhir ditentukan melalui voting mayoritas dari seluruh pohon. Pada penelitian ini, Random Forest dikonfigurasi dengan 200 pohon keputusan ($n_estimators=200$) tanpa pembatasan kedalaman maksimum, mengikuti konfigurasi umum yang digunakan pada studi klasifikasi status gizi sebelumnya [10], [18].

Support Vector Machine (SVM) bekerja dengan mencari hyperplane pemisah optimal yang memaksimalkan margin antarkelas pada ruang fitur. Pada penelitian ini digunakan kernel Radial Basis Function (RBF) untuk menangani hubungan non-linear antara fitur antropometri dan kelas status gizi, dengan parameter $probability=True$ agar model dapat menghasilkan estimasi probabilitas kelas yang dibutuhkan pada tahap stacking. Al-Nafjan dkk. [33] mencatat bahwa algoritma berbasis pohon seperti Random Forest dan XGBoost paling banyak digunakan untuk data klinis terstruktur karena kemampuannya menangkap hubungan non-linear antarfitur. XGBoost merupakan algoritma ensemble berbasis gradient boosting yang membangun pohon keputusan secara sekuensial, di mana setiap pohon baru dilatih untuk memperbaiki kesalahan residual dari pohon-pohon sebelumnya, dilengkapi dengan regularisasi L1 dan L2 untuk mencegah overfitting. Pada penelitian ini, XGBoost dikonfigurasi dengan 200 estimator dan fungsi log-loss multikelas (mlogloss) sebagai metrik evaluasi internal.

Mekanisme stacking ensemble pada penelitian ini bekerja dalam dua tingkat (level). Pada tingkat pertama, ketiga base learner (Random Forest, SVM, dan XGBoost) dilatih secara terpisah menggunakan skema 5-fold cross-validation terhadap data latih, menghasilkan prediksi probabilitas out-of-fold untuk setiap kelas dari masing-masing base learner. Prediksi probabilitas tersebut kemudian digunakan sebagai fitur input bagi meta-learner pada tingkat kedua, yaitu Logistic Regression, yang mempelajari bobot optimal untuk mengombinasikan prediksi ketiga base learner tersebut menjadi satu prediksi akhir. Skema cross-validation internal ini penting untuk mencegah meta-learner mempelajari pola yang terlalu optimistik akibat base learner yang dievaluasi pada data yang sama dengan data pelatihannya sendiri (data leakage).

Secara matematis, mekanisme stacking ensemble pada penelitian ini dapat dirumuskan sebagai berikut. Untuk setiap data x , ketiga base learner (Random Forest, SVM, dan XGBoost) menghasilkan estimasi probabilitas kelas yang kemudian digabungkan menjadi satu vektor meta-fitur $Z(x)$ sebagaimana Persamaan (2).

$$Z(x) = [P_{RF}(y|x), P_{SVM}(y|x), P_{XGB}(y|x)] \quad (2)$$

Vektor meta-fitur $Z(x)$ tersebut kemudian menjadi input bagi meta-learner Logistic Regression, yang menghasilkan probabilitas akhir untuk setiap kelas k melalui fungsi softmax sebagaimana Persamaan (3), dengan w_k dan b_k masing-masing menyatakan bobot dan bias hasil pelatihan meta-learner untuk kelas k .

$$P(y = k|Z(x)) = \frac{\exp(w_k \cdot Z(x) + b_k)}{\sum_j \exp(w_j \cdot Z(x) + b_j)} \quad (3)$$

Prediksi akhir model stacking ensemble $\hat{y}$ untuk data x kemudian ditentukan dengan memilih kelas yang memiliki probabilitas tertinggi, yaitu $\hat{y} = \operatorname{argmax}_k P(y=k|Z(x))$.

E. Metrik Evaluasi

Performa model dievaluasi menggunakan metrik akurasi, presisi makro, recall makro, dan F1-score makro, serta divisualisasikan melalui confusion matrix untuk melihat pola kesalahan klasifikasi antarkelas. Penggunaan metrik makro dipilih karena distribusi kelas pada kedua dataset tidak seimbang, sehingga metrik makro memberikan bobot yang setara terhadap setiap kelas tanpa didominasi oleh kelas mayoritas.

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Presisi} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score = 2 \times \frac{Presisi \times Recall}{Presisi + Recall} \quad (7)$$

dengan TP, TN, FP, dan FN masing-masing menyatakan jumlah true positive, true negative, false positive, dan false negative pada setiap kelas, sebagaimana dihitung menggunakan Persamaan (4)-(7) di atas. Untuk metrik makro, nilai presisi, recall, dan F1-score dihitung terlebih dahulu per kelas, kemudian dirata-rata secara sederhana (tanpa pembobotan) terhadap seluruh kelas.

F. Lingkungan Implementasi

Seluruh tahapan penelitian, mulai dari pembersihan data, praproses, pelatihan model, hingga evaluasi, diimplementasikan menggunakan bahasa pemrograman Python di dalam lingkungan notebook interaktif (*Jupyter Notebook*). Pustaka utama yang digunakan meliputi pandas dan NumPy untuk manipulasi data, scikit-learn untuk praproses dan implementasi Random Forest, SVM, Logistic Regression, serta kerangka kerja stacking, dan pustaka XGBoost untuk implementasi algoritma gradient boosting. Visualisasi hasil, termasuk grafik perbandingan performa dan confusion matrix, dibuat menggunakan pustaka Matplotlib dan Seaborn. Seluruh proses eksperimen menggunakan nilai random state yang tetap (random_state=42) pada setiap tahapan yang melibatkan unsur acak, seperti pembagian data dan inisialisasi model, guna menjamin reproduktifitas hasil penelitian.

3. HASIL DAN PEMBAHASAN

A. Karakteristik Data

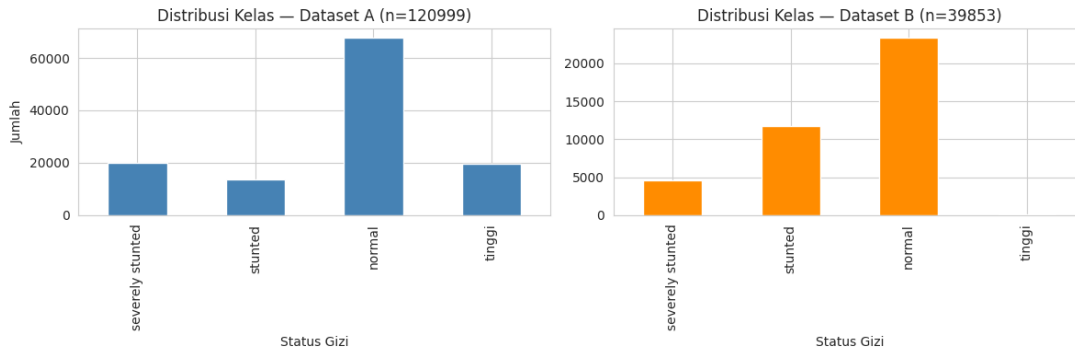
Tabel 1 menunjukkan statistik deskriptif fitur numerik pada kedua dataset setelah proses pembersihan. Dataset A memiliki rentang usia 0–60 bulan dengan rata-rata tinggi badan 88,66 cm, sedangkan dataset B memiliki rentang usia yang sama dengan rata-rata tinggi badan 86,04 cm. Distribusi kelas status gizi pada dataset A relatif lebih seimbang dibandingkan dataset B, di mana kelas normal mendominasi dataset B (23.387 dari 39.853 data), sementara kelas tinggi pada dataset B hanya berjumlah 53 data, menunjukkan ketidakseimbangan kelas yang signifikan pada dataset lapangan tersebut.

Tabel 1. Statistik Deskriptif Fitur Numerik

Statistik	Usia (bulan) A	Tinggi (cm) A	Usia (bulan) B	Tinggi (cm) B
N	120.999	120.999	39.853	39.853
Rata-rata	30,17	88,66	33,08	86,04
Std. Deviasi	17,58	17,30	15,81	11,80
Minimum	0	40,01	0	44,00
Maksimum	60	128,00	60	120,00

Sumber: Hasil pengolahan data, 2026

Untuk memperjelas gambaran distribusi kelas pada Tabel 1, Gambar 2 berikut menyajikan visualisasi perbandingan jumlah data pada setiap kelas status gizi di kedua dataset.



Gambar 2. Distribusi Kelas Status Gizi pada Dataset A dan Dataset B

Gambar 2 memperkuat gambaran pada Tabel 1, yaitu distribusi kelas pada dataset A relatif lebih merata di antara keempat kategori dibandingkan dataset B, di mana kelas normal mendominasi secara signifikan dan kelas tinggi hampir tidak terlihat karena jumlahnya yang sangat sedikit (53 data). Perbedaan distribusi kelas inilah yang turut berkontribusi terhadap penurunan performa pada skema validasi lintas dataset, sebagaimana dibahas pada bagian selanjutnya.

B. Perbandingan Performa Model Dalam-Dataset

Tabel 2 menunjukkan hasil evaluasi model pada skema pengujian dalam-dataset menggunakan dataset A (data uji sebanyak 24.200 data). Model stacking ensemble memperoleh performa tertinggi pada seluruh metrik, dengan akurasi 99,94% dan F1-score makro 99,90%, sedikit mengungguli Random Forest yang juga menunjukkan performa sangat tinggi (akurasi 99,93%). XGBoost memperoleh akurasi 99,12%, sementara SVM menghasilkan performa relatif terendah di antara keempat model dengan akurasi 98,37%, meskipun nilai ini masih tergolong sangat tinggi secara absolut.

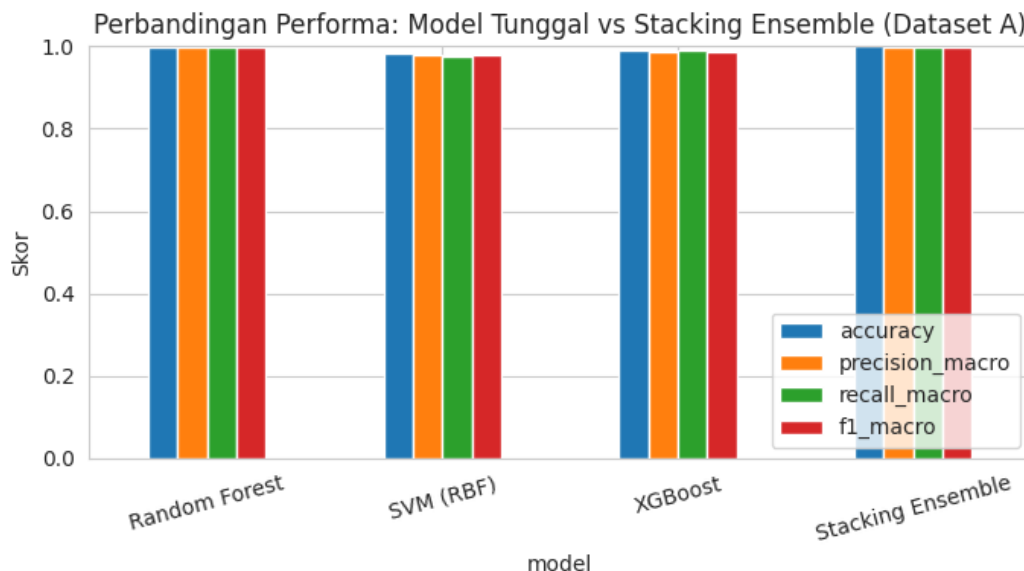
Tabel 2. Perbandingan Performa Model pada Pengujian Dalam-Dataset (Dataset A)

Model	Akurasi	Presisi Makro	Recall Makro	F1 Makro
Random Forest	0,9993	0,9991	0,9987	0,9989
SVM (RBF)	0,9837	0,9808	0,9743	0,9775
XGBoost	0,9912	0,9862	0,9890	0,9875
Stacking Ensemble	0,9994	0,9991	0,9989	0,9990

Sumber: Hasil pengolahan data, 2026

Peningkatan performa stacking ensemble terhadap model terbaiknya (Random Forest) tergolong marginal, yaitu sekitar 0,01 poin persentase pada akurasi. Temuan ini konsisten dengan beberapa penelitian yang menunjukkan bahwa penerapan stacking ensemble pada dataset dengan pola yang relatif mudah dipelajari cenderung memberikan peningkatan performa yang tidak signifikan dibandingkan base learner terbaiknya, meskipun tetap unggul dibandingkan model dengan performa terendah [30]. Nilai akurasi yang sangat tinggi (mendekati 100%) pada seluruh model mengindikasikan bahwa pola hubungan antara fitur antropometri dan label status gizi pada dataset A relatif mudah dipelajari, kemungkinan karena label pada dataset tersebut dihasilkan melalui aturan ambang batas yang konsisten, sebagaimana lazim ditemukan pada dataset publik hasil kurasi atau simulasi.

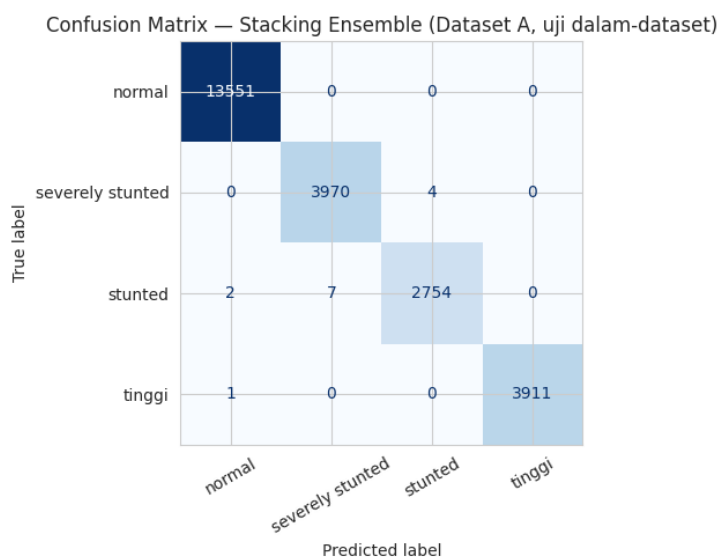
Perbandingan keempat model tersebut secara visual disajikan pada Gambar 3 berikut.



Gambar 3. Perbandingan Performa Model Tunggal vs Stacking Ensemble (Dataset A)

Gambar 3 memperlihatkan bahwa keempat batang metrik (akurasi, presisi makro, recall makro, dan F1 makro) pada model Random Forest dan Stacking Ensemble berada pada posisi yang hampir setara dan paling tinggi di antara keempat model, sementara batang pada model SVM tampak sedikit lebih rendah dibandingkan ketiga model lainnya. Visualisasi ini menegaskan bahwa selisih performa antarmodel pada skema dalam-dataset tergolong tipis secara keseluruhan.

Untuk melihat lebih rinci pola kesalahan klasifikasi model stacking ensemble per kelas, Gambar 4 menyajikan confusion matrix hasil pengujian model tersebut pada data uji dataset A.



Gambar 4. Confusion Matrix Model Stacking Ensemble pada Pengujian Dalam-Dataset (Dataset A)

Gambar 4 menunjukkan bahwa mayoritas prediksi model stacking ensemble terkonsentrasi pada diagonal utama confusion matrix, mengindikasikan tingkat kesalahan klasifikasi yang sangat rendah pada skema dalam-dataset. Kesalahan yang tersisa terutama terjadi pada pasangan kelas yang berdekatan

secara ambang batas z-score, yaitu antara kelas stunted dan severely stunted, sebagaimana dibahas lebih lanjut pada bagian F.

C. Hasil Validasi Lintas Dataset

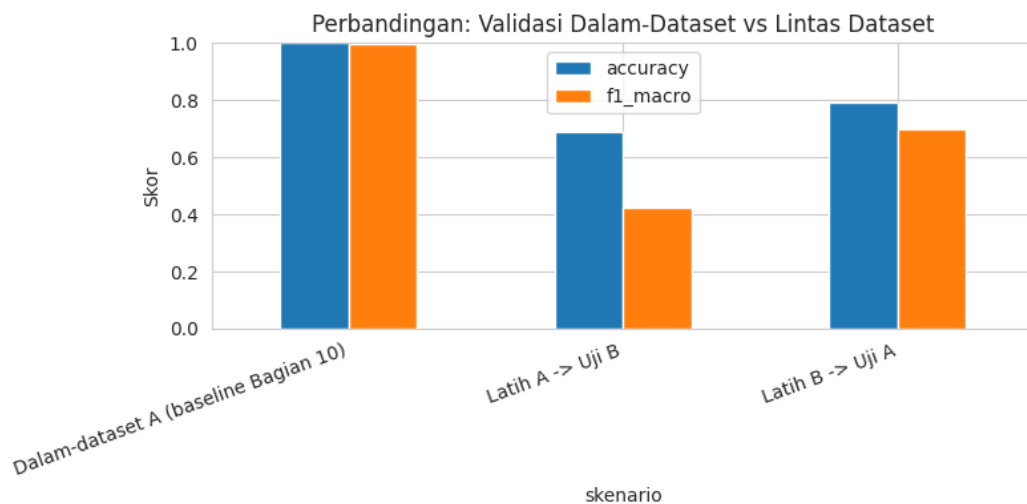
Tabel 3 menyajikan hasil validasi lintas dataset dibandingkan dengan hasil pengujian dalam-dataset sebagai acuan. Hasil menunjukkan penurunan performa yang sangat signifikan pada kedua arah pengujian lintas dataset. Ketika model stacking ensemble dilatih pada dataset A dan diuji pada dataset B, akurasi menurun tajam dari 99,94% (dalam-dataset) menjadi 68,80%, dengan F1-score makro menurun lebih drastis lagi dari 99,90% menjadi 42,38%. Pada arah sebaliknya, yaitu model dilatih pada dataset B dan diuji pada dataset A, akurasi mencapai 79,12% dengan F1-score makro 69,78%, lebih baik dibandingkan arah A ke B, namun tetap jauh di bawah performa dalam-dataset.

Tabel 3. Perbandingan Performa Dalam-Dataset dan Lintas Dataset (Model Stacking Ensemble)

Skenario Pengujian	Akurasi	Presisi Makro	Recall Makro	F1 Makro
Dalam-dataset A (baseline)	0,9994	0,9991	0,9989	0,9990
Latih A → Uji B	0,6880	0,4312	0,4859	0,4238
Latih B → Uji A	0,7912	0,7883	0,6801	0,6978

Sumber: Hasil pengolahan data, 2026

Perbedaan performa antara skema dalam-dataset dan lintas dataset pada Tabel 3 divisualisasikan pada Gambar 5 berikut agar besarnya penurunan performa dapat diamati secara langsung.



Gambar 5. Perbandingan Akurasi dan F1-Score Makro: Dalam-Dataset vs Lintas Dataset

Gambar 5 memperlihatkan secara visual bahwa batang akurasi dan F1-score makro pada skenario dalam-dataset jauh lebih tinggi dibandingkan kedua skenario lintas dataset, dengan penurunan F1-score makro yang tampak lebih tajam dibandingkan penurunan akurasi. Perbedaan ini konsisten dengan sifat metrik F1-score makro yang memberi bobot setara pada setiap kelas, sehingga lebih sensitif terhadap kegagalan model dalam mengklasifikasikan kelas minoritas secara benar dibandingkan metrik akurasi yang cenderung didominasi oleh kelas mayoritas.

Penurunan performa yang lebih tajam pada arah Latih A → Uji B dibandingkan arah sebaliknya kemungkinan disebabkan oleh karakteristik dataset A yang bersifat lebih homogen dan berpola sederhana (sebagaimana terindikasi dari performa dalam-dataset yang mendekati sempurna), sehingga model yang dilatih pada dataset tersebut kurang mampu menangkap variasi pola yang lebih kompleks pada data lapangan riil (dataset B). Sebaliknya, model yang dilatih pada dataset B, yang mengandung

variasi data yang lebih realistis karena berasal dari pengukuran lapangan langsung, tampak memiliki kemampuan generalisasi yang relatif lebih baik ketika diuji pada dataset A.

Temuan mengenai kesenjangan generalisasi (generalization gap) ini sejalan dengan fenomena domain shift yang telah banyak dibahas pada domain machine learning lain, di mana performa model yang sangat tinggi pada pengujian dalam-domain seringkali tidak berbanding lurus dengan kemampuan generalisasinya terhadap domain atau sumber data yang berbeda. Pada konteks klasifikasi status gizi balita, hasil ini mengonfirmasi bahwa evaluasi model hanya dengan skema latih-uji tunggal pada satu dataset, sebagaimana dilakukan pada sebagian besar penelitian terdahulu [5]–[19], berisiko memberikan gambaran performa yang terlalu optimistis dan tidak merepresentasikan performa riil model ketika diterapkan pada populasi yang berbeda, misalnya pada wilayah dengan karakteristik demografis dan pola pengukuran yang berbeda dari data pelatihan.

D. Perbandingan dengan Penelitian Terdahulu

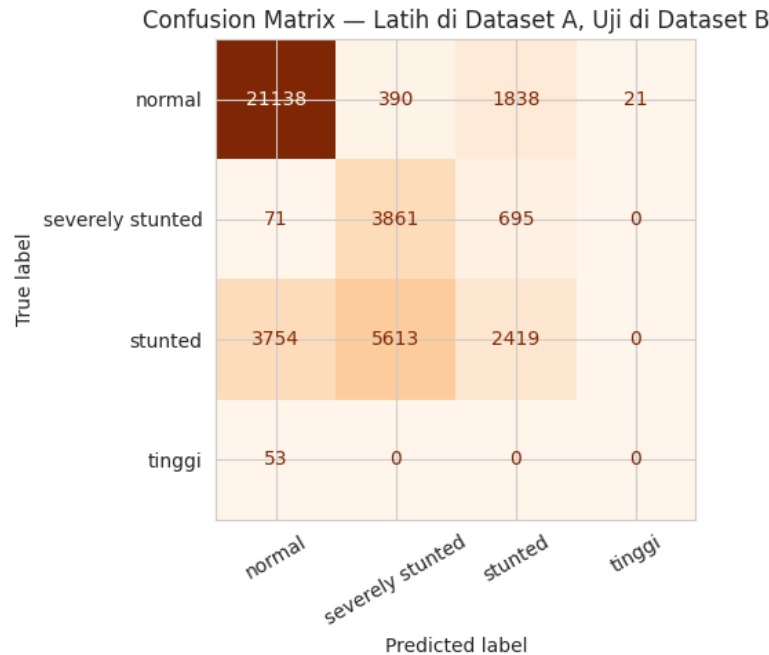
Dibandingkan dengan penelitian Elim dan Utami [10] yang memperoleh akurasi sempurna (1,00) menggunakan Random Forest dengan optimasi grid search pada data stunting berskala kecil, maupun penelitian Fadil dkk. [11] yang menerapkan SVM dengan SMOTE untuk prediksi stunting multikelas di Kabupaten Sumedang, performa dalam-dataset pada penelitian ini menunjukkan pola yang sebanding, yaitu sangat tinggi ketika model diuji pada dataset yang sama dengan sumber pelatihannya. Perbedaan performa antar-studi ini kemungkinan besar disebabkan oleh karakteristik dataset yang berbeda, dataset publik berskala besar seperti dataset A pada penelitian ini cenderung memiliki pola yang lebih konsisten dibandingkan data lapangan berskala kecil yang digunakan pada penelitian-penelitian tersebut. Hal ini semakin memperkuat pentingnya validasi lintas dataset sebagai bagian dari evaluasi model, karena performa yang dilaporkan pada satu dataset tidak dapat digeneralisasikan secara langsung ke dataset atau konteks lain, sebagaimana ditunjukkan pula pada penelitian Bitew dkk. [6] yang menemukan variasi performa algoritma berbasis pohon pada populasi Ethiopia.

Perbandingan lebih lanjut dengan penelitian Siramaneerat dkk. [15] yang menganalisis determinan stunting secara multilevel pada wilayah urban dan rural di Indonesia menunjukkan pola yang serupa dengan temuan penelitian ini, yaitu performa model pada data lapangan skala kecil dan spesifik wilayah cenderung lebih rendah dan lebih bervariasi dibandingkan performa pada dataset berskala besar. Hal ini juga sejalan dengan hasil validasi lintas dataset pada penelitian ini, di mana performa terbaik justru diperoleh ketika model dilatih pada dataset lapangan (dataset B, Jeneponto) dan diuji pada dataset skala besar (dataset A), dibandingkan arah sebaliknya. Kondisi ini mengindikasikan bahwa data lapangan riil, meskipun berjumlah lebih sedikit, dapat memuat pola variasi alami yang lebih representatif dibandingkan data hasil kurasi atau simulasi berskala besar, sehingga model yang dilatih pada data tersebut cenderung lebih tahan terhadap perubahan distribusi data (distributional shift).

Sementara itu, dibandingkan dengan penelitian Anku dan Duah [5] di Ghana maupun Zemariam dkk. [7] di Ethiopia yang turut melaporkan variasi performa algoritma machine learning antarwilayah dalam satu negara, penelitian ini memberikan bukti tambahan bahwa variasi performa tersebut dapat terjadi bahkan ketika model diuji secara eksplisit lintas sumber dataset yang berbeda karakteristik pengumpulan datanya, bukan sekadar lintas wilayah dalam skema pengumpulan data yang seragam. Hal ini memperkuat argumen bahwa laporan akurasi tunggal pada satu skema evaluasi tidak cukup untuk menyimpulkan kesiapan suatu model machine learning untuk diterapkan secara luas dalam mendukung program penurunan stunting nasional maupun daerah [2], [3].

E. Analisis Pola Kesalahan Klasifikasi

Untuk memahami sumber penurunan performa pada skenario Latih A → Uji B secara lebih rinci, Gambar 6 menyajikan confusion matrix hasil pengujian tersebut per kelas.



Gambar 6. Confusion Matrix Model Stacking Ensemble: Latih pada Dataset A, Uji pada Dataset B

Analisis lebih lanjut terhadap confusion matrix model stacking ensemble pada skema dalam-dataset menunjukkan bahwa kesalahan klasifikasi yang tersisa sebagian besar terjadi antara kelas stunted dan severely stunted, yaitu dua kelas yang berbatasan langsung pada ambang z-score -3. Pola ini sejalan dengan karakteristik data antropometri secara umum, di mana nilai tinggi badan pada balita yang berada tepat di sekitar ambang batas klasifikasi cenderung sulit dibedakan secara tegas, bahkan oleh tenaga kesehatan terlatih sekalipun. Pola kesalahan serupa juga ditemukan pada confusion matrix hasil validasi lintas dataset (Latih A → Uji B), dengan tambahan kecenderungan model salah mengklasifikasikan sejumlah besar data kelas stunted dan severely stunted pada dataset B ke dalam kelas normal. Hal ini mengindikasikan bahwa model yang dilatih pada dataset A cenderung terlalu percaya diri (overconfident) dalam memprediksi kelas normal ketika dihadapkan pada distribusi data yang berbeda, kemungkinan karena proporsi kelas normal yang dominan pada dataset B (23.387 dari 39.853 data, atau sekitar 58,7%).

Dari sisi implikasi praktis, temuan mengenai kesenjangan generalisasi pada penelitian ini memiliki relevansi penting bagi pengembangan sistem deteksi dini stunting berbasis machine learning yang akan diterapkan di lapangan, misalnya pada aplikasi pemantauan gizi di Posyandu atau Puskesmas. Model yang hanya dievaluasi dan menunjukkan performa sangat tinggi pada satu sumber data, sebagaimana banyak dilaporkan pada penelitian terdahulu [10]–[19], berisiko memberikan hasil yang kurang akurat ketika diterapkan pada populasi dengan karakteristik demografis, pola pengukuran, atau kualitas data yang berbeda dari data pelatihannya. Oleh karena itu, sebelum model semacam ini diimplementasikan secara luas untuk mendukung intervensi gizi masyarakat sebagaimana diamanatkan dalam kerangka investasi penurunan stunting [2], evaluasi lintas populasi atau lintas wilayah perlu menjadi bagian standar dari proses validasi model, bukan sekadar pelengkap opsional.

F. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu menjadi perhatian pada penelitian selanjutnya. Pertama, label status gizi pada dataset B merupakan hasil turunan dari nilai z-score menggunakan ambang WHO [1], bukan label asli yang disediakan langsung oleh sumber data, sehingga berpotensi memuat kesalahan klasifikasi pada nilai z-score yang berada tepat di batas ambang. Kedua,

pemetaan kolom jenis kelamin (Gender) pada dataset B dari nilai numerik 0/1 ke kategori laki-laki/perempuan dilakukan berdasarkan asumsi konvensi umum, sehingga perlu diverifikasi lebih lanjut terhadap dokumentasi resmi sumber data. Ketiga, kelas tinggi pada dataset B memiliki jumlah data yang sangat sedikit (53 data), yang berpotensi memengaruhi keandalan estimasi metrik makro pada kelas tersebut. Keempat, penelitian ini hanya menggunakan fitur antropometri dasar tanpa melibatkan data sosioekonomi, sehingga kemampuan model dalam menangkap faktor risiko stunting yang bersifat multidimensi masih terbatas.

Secara keseluruhan, temuan penelitian ini memberikan dua kontribusi utama terhadap literatur klasifikasi status gizi balita berbasis machine learning. Pertama, penelitian ini mengonfirmasi bahwa penerapan *stacking ensemble* mampu memberikan performa yang setara atau sedikit lebih baik dibandingkan *base learner* terbaiknya pada skema evaluasi dalam-dataset, sejalan dengan temuan tinjauan sistematis Mahajan dkk. [31]. Kedua, dan yang lebih penting, penelitian ini menunjukkan secara empiris bahwa metrik performa yang sangat tinggi pada skema evaluasi konvensional (single train-test split) tidak dapat dijadikan satu-satunya acuan dalam menilai kesiapan model untuk diterapkan pada konteks yang lebih luas, sebuah aspek yang jarang dieksplisitkan pada studi-studi klasifikasi stunting sebelumnya di Indonesia maupun negara berkembang lainnya [5]–[24].

4. KESIMPULAN

Model *stacking ensemble* yang mengombinasikan Random Forest, SVM, dan XGBoost dengan meta-learner Logistic Regression berhasil memberikan performa klasifikasi status gizi balita terbaik pada skema pengujian dalam-dataset, dengan akurasi 99,94% dan F1-score makro 99,90%, meskipun peningkatannya terhadap model tunggal terbaik (Random Forest) tergolong marginal. Validasi lintas dataset menunjukkan penurunan performa yang signifikan, dengan akurasi menurun menjadi 68,80% (Latih A → Uji B) dan 79,12% (Latih B → Uji A), mengindikasikan adanya kesenjangan generalisasi (*generalization gap*) yang substansial antara performa dalam-dataset dan performa lintas sumber data.

Temuan ini menegaskan bahwa evaluasi model klasifikasi status gizi balita tidak dapat hanya mengandalkan skema latihan-uji tunggal pada satu dataset, karena performa yang tampak sangat tinggi belum tentu merepresentasikan kemampuan model saat diterapkan pada populasi atau wilayah yang berbeda karakteristik datanya. Penelitian selanjutnya disarankan untuk menggabungkan data dari berbagai sumber dan wilayah dalam proses pelatihan (*multi-source training*), menerapkan teknik *domain adaptation* untuk mengurangi kesenjangan generalisasi, memperkaya fitur dengan data sosioekonomi dan lingkungan, serta memverifikasi label dan metadata dataset secara langsung kepada penyedia data guna meningkatkan keandalan hasil penelitian.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Potensi Utama atas dukungan yang diberikan selama proses penelitian ini. Penulis juga berterima kasih kepada penyedia dataset publik pada *Kaggle* dan *Mendeley Data* yang telah memungkinkan penelitian ini dilaksanakan.

REFERENSI

- [1] Mulyani, A. T., Khairinisa, M. A., Khatib, A., & Chaerunisaa, A. Y. (2025). Understanding stunting: Impact, causes, and strategy to accelerate stunting reduction—A narrative review. *Nutrients*, 17(9), 1493. <https://doi.org/10.3390/nu17091493>
- [2] Komarulzaman, A., Smits, J., de Jong, E., Karim, M. R., & Mubasyira, M. (2023). Achieving zero stunting: A sustainable development goal interlinkage approach at district level. *Sustainability*, 15(11), 8890. <https://doi.org/10.3390/su15118890>

- [3] Novalina, N., Tarigan, I. A. A., Kameela, F. K., & Rizkinia, M. (2025). Benchmarking machine learning algorithm for stunting risk prediction in Indonesia. *Bulletin of Electrical Engineering and Informatics*, 14(3), 2252–2263.
- [4] Khan, M. N. A., & Yunus, R. M. (2023). A hybrid ensemble approach to accelerate the classification accuracy for predicting malnutrition among under-five children in sub-Saharan African countries. *Nutrition*, 108, 111947. <https://doi.org/10.1016/j.nut.2022.111947>
- [5] Anku, E. K., & Duah, H. O. (2024). Predicting and identifying factors associated with undernutrition among children under five years in Ghana using machine learning algorithms. *PLOS ONE*, 19(2), e0296625. <https://doi.org/10.1371/journal.pone.0296625>
- [6] Bitew, F. H., Sparks, C. S., & Nyarko, S. H. (2022). Machine learning algorithms for predicting undernutrition among under-five children in Ethiopia. *Public Health Nutrition*, 25(2), 269–280.
- [7] Zemariam, A. B., Abate, B. B., Alamaw, A. W., Lake, E. S., Yilak, G., Ayele, M., Tilahun, B. D., & Ngusie, H. S. (2025). Prediction of stunting and its socioeconomic determinants among adolescent girls in Ethiopia using machine learning algorithms. *PLOS ONE*, 20(1), e0316452. <https://doi.org/10.1371/journal.pone.0316452>
- [8] Rao, B., Rashid, M., Hasan, M. G., & Thunga, G. (2025). Machine learning in predicting child malnutrition: A meta-analysis of demographic and health surveys data. *International Journal of Environmental Research and Public Health*, 22(3), 449. <https://doi.org/10.3390/ijerph22030449>
- [9] Puspita, D. M., Santoso, K. I., Triyono, A., Supriyadi, E., & Agus, S. (2025). Optimization performance of extreme gradient boosting and random forest for child stunting classification based on economic factors. *Jurnal Teknik Informatika (JUTIF)*, 7(3), 2967–2977.
- [10] Elim, M. I., & Utami, E. (2025). Performance comparison of child stunting prediction: Support vector machine vs random forest with grid search optimization. *Jurnal Teknik Informatika (JUTIF)*, 6(5), 5305–5319. <https://doi.org/10.52436/1.jutif.2025.6.5.5285>
- [11] Fadil, I., Surya Manggala, R., Firmansyah, E., & Helmiawan, M. A. (2025). Performance optimization of support vector machine with SMOTE for multiclass stunting prediction in Sumedang District, Indonesia. *Jurnal Teknik Informatika (JUTIF)*, 6(4), 2917–2928. <https://doi.org/10.52436/1.jutif.2025.6.4.4843>
- [12] Hasdyna, N., Dinata, R. K., Rahmi, & Fajri, T. I. (2024). Hybrid machine learning for stunting prevalence: A novel comprehensive approach to its classification, prediction, and clustering optimization in Aceh, Indonesia. *Informatics*, 11(4), 89. <https://doi.org/10.3390/informatics11040089>
- [13] Hendy, A., Ibrahim, R. K., Abdelaliem, S. M. F., Zaher, A., Alkubati, S. A., Abd El-kader, R. G., & Hendy, A. (2025). Supervised machine learning for classification and prediction of stunting among under-five Egyptian children. *BMC Pediatrics*, 25, 681. <https://doi.org/10.1186/s12887-025-06138-x>
- [14] Azriani, D., Agustian, D., Zuhairini, Y., Yulita, I. N., & Dhamayanti, M. (2025). Prediction models for stunting at 2-years-old from Indonesian newborn population. *BMC Pediatrics*, 25(1). <https://doi.org/10.1186/s12887-025-06096-4>
- [15] Siramaneerat, I., Astutik, E., Agushybana, F., Bhumkittipich, P., & Lamprom, W. (2024). Examining determinants of stunting in urban and rural Indonesian: A multilevel analysis using the population-based Indonesian family life survey (IFLS). *BMC Public Health*, 24(1), 1–13. <https://doi.org/10.1186/s12889-024-18824-z>
- [16] Ndagijimana, S., Kabano, I. H., Masabo, E., & Ntaganda, J. M. (2023). Prediction of stunting among under-5 children in Rwanda using machine learning techniques. *Journal of Preventive Medicine & Public Health*, 56, 41–49.

- [17] Ndagijimana, S., Kabano, I. H., Masabo, E., & Ntaganda, J. M. (2024). Predicting stunting in Rwanda using artificial neural networks: A demographic health survey 2020 analysis. *F1000Research*, 13, 128.
- [18] Ndagijimana, A., Elfving, K., Umubyeyi, A., & Lind, T. (2025). Identification of amendable risk factors for childhood stunting at individual, household and community levels in Northern Province, Rwanda: A cross-sectional population-based study. *BMC Public Health*, 25.
- [19] Sur, K., Kundu, T., Dhote, S., & Singh, S. K. (2025). Unveiling predictive factors for household-level stunting in India: A machine learning approach using NFHS-5 and satellite-driven data. *Nutrition*, 132, 112674. <https://doi.org/10.1016/j.nut.2024.112674>
- [20] Fogarty, B., García-Martínez, A., Chawla, N. V., & Serván-Mori, E. (2025). Social and economic predictors of under-five stunting in Mexico: A comprehensive approach through the XGB model. *Journal of Global Health*, 15.
- [21] Begashaw, G. B., Zewotir, T., & Fenta, H. M. (2025). A deep learning approach for classifying and predicting children's nutritional status in Ethiopia using LSTM-FC neural networks. *BioData Mining*, 18(1), 11.
- [22] Chibuye, M., Mende, D. R., Spijker, R., Simuyandi, M., Luchen, C. C., Bosomprah, S., Chilengi, R., Schultz, C., & Harris, V. C. (2024). Systematic review of associations between gut microbiome composition and stunting in under-five children. *npj Biofilms and Microbiomes*, 10(1), 41. <https://doi.org/10.1038/s41522-024-00517-5>
- [23] Akuze, J., Otieno, G. P., Amare, S. Y., Ngatia, B., Wanduru, P., Kirakoya-Samadoulougou, F., Kananura, R. M., Amouzou, A., Estifanos, A. S., & Ohuma, E. O. (2025). Data science and artificial intelligence for maternal, newborn and child health: Scoping review and thematic analysis. *BMC Public Health*, 25.
- [24] Richard, W., Thangata, P., Mkandawire, B., & Amoah, N. (2025). Advancing predictive analytics in child malnutrition: Machine, ensemble and deep learning models with balanced class distribution for early detection of stunting and wasting. *Human Nutrition & Metabolism*, 42, 200340. <https://doi.org/10.1016/j.hnm.2025.200340>
- [25] Lailan, M., Azzawagama, A., Muslih, A., Zukhruf, G., & Sirnopati, R. (2026). Integrating education-based interventions and machine learning for stunting prevention: A case study in East Lombok, Indonesia. *Dialogues in Health*, 8, 100264. <https://doi.org/10.1016/j.dialog.2025.100264>
- [26] Islam, M., Rahman, M. J., Chaki, D., & Biswas, M. (2022). Application of machine learning based algorithm for prediction of malnutrition among women in Bangladesh. *International Journal of Cognitive Computing in Engineering*, 3, 46–57. <https://doi.org/10.1016/j.ijcce.2022.01.001>
- [27] Ankalaki, S., G Biradar, V., Naik P, K. K., & S. Hukkeri, G. (2024). A deep learning approach for malnutrition detection. *International Journal of Online and Biomedical Engineering (iJOE)*, 20(6), 116–133.
- [28] Biradar, V. G., & Naik, K. K. (2024). Detection of malnutrition in children using deep learning model. In T. Senjyu, C. So-In, & A. Joshi (Eds.), *Smart Trends in Computing and Communications (SmartCom 2024), Lecture Notes in Networks and Systems (Vol. 946)*. Springer. https://doi.org/10.1007/978-981-97-1323-3_4
- [29] Ramón, A., Torres, A. M., Milara, J., Mateo, J., & Cortés Castell, E. (2022). eXtreme gradient boosting-based method to classify patients with COVID-19. *Journal of Investigative Medicine*, 70(7), 1472–1480. <https://doi.org/10.1136/jim-2021-002205>
- [30] Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with stacking ensemble learner for predicting emergency readmission of heart-disease patients. *Decision Analytics Journal*, 7, 100242. <https://doi.org/10.1016/j.dajour.2023.100242>
- [31] Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble learning for disease prediction: A review. *Healthcare*, 11(12), 1808. <https://doi.org/10.3390/healthcare11121808>

- [32] Sifat, I. K., Kibria, M. K., & Popovic, T. (2024). Optimizing hypertension prediction using ensemble learning approaches. *PLOS ONE*, 19(12), e0315865. <https://doi.org/10.1371/journal.pone.0315865>
- [33] Al-Nafjan, A., Aljuhani, A., Alshebel, A., Alharbi, A., & Alshehri, A. (2025). Artificial intelligence in predictive healthcare: A systematic review. *Journal of Clinical Medicine*, 14(19), 6752. <https://doi.org/10.3390/jcm14196752>
- [34] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science*, 17(11), e70056. <https://doi.org/10.1111/cts.70056>
- [35] Im, C., Kim, W., & Kim, H. (2025). Explainable machine learning for heat-related illness prediction: An XGBoost–SHAP approach using Korean meteorological data. *Bioengineering*, 12(11), 1276. <https://doi.org/10.3390/bioengineering12111276>