TEKNOMATIKA
JURNAL INFORMATIKA DAN KOMPUTER
Vol. 17, No. 2, Desember 2024, pp. 38-47
ISSN: 1979-7656 e-ISSN: 3031-0865
DOI: 10.30989/teknomatika.v17i2.1530

Klasifikasi Data Tak Seimbang Menggunakan Algoritma Random Forest dengan SMOTE dan SMOTE-ENN (Studi Kasus pada Data Stunting)

Anju Fauziah[a,1,*], Julan Hernadi[a,2]
[a]Matematika Universitas Ahmad Dahlan, Jln. Ringroad Selatan, Yogyakarta
[1]anju2000015021@webmail.uad.ac.id, [2]julan.hernadi@math.uad.ac.id
[*]Penulis koresponden

Diterima	Direvisi	Disetujui	Dipublikasikan
13/12/2024	19/12/2024	20/12/2024	30/12/2024

ABSTRACT

The random forest algorithm is one of the widely used machine learning classification methods because it has the advantage of reducing the risk of overfitting while improving general prediction performance. However, for data with unbalanced classes, this algorithm lacks to achieve its best performance, particularly in predicting data in the minority class. As a result, this article proposes two resampling approaches to balance the data: the Synthetic Minority Oversampling Technique (SMOTE) and the Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN). For the data classification technique, the random forest algorithm is applied to the original data, then to the resampling results using both SMOTE as well as SMOTE-ENN. The case study was applied to stunting data consisting of 421 cases in the majority class and 79 in the minority class. An accuracy of 89% was obtained on the original data, 90% on the resampled data with SMOTE-ENN, and 91% on the resampled data with SMOTE. The best accuracy was obtained using resampling technique with SMOTE, however it was not particularly significant.

KEYWORDS

Classification
Imbalanced Class
Random Forest
SMOTE
SMOTE-ENN

ABSTRAK

Algoritma random forest merupakan salah satu metode klasifikasi pembelajaran mesin yang banyak digunakan karena memiliki keunggulan dalam mengurangi resiko overfitting sekaligus meningkatkan kinerja prediksi secara umum. Namun untuk data dengan kelas tidak seimbang, algoritma ini tidak mampu mencapai performa maksimal khususnya dalam memprediksi data pada kelas minoritas. Untuk itu artikel ini menawarkan dua metode resampling untuk menyeimbangkan data, yaitu Synthetic Minority Oversampling Technique (SMOTE) dan Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN). Untuk klasifikasi data diterapkan algoritma random forest terhadap data asli dan hasil resampling baik menggunakan SMOTE maupun SMOTE-ENN. Studi kasus diterapkan pada data stunting yang berjumlah 421 pada kelas mayoritas dan 79 pada kelas minoritas. Diperoleh akurasi 89% pada data asli, 90% pada data hasil resampling dengan SMOTE-ENN, dan 91% pada data resampling dengan SMOTE. Walaupun tidak terlalu signifikan, teknik resampling dengan SMOTE memberikan akurasi terbaik.

KATA KUNCI
Klasifikasi
Imbalanced Class
Random Forest
SMOTE
SMOTE-ENN

1 PENDAHULUAN

Sunting adalah kondisi gangguan pertumbuhan dan perkembangan pada balita akibat kekurangan gizi yang berlangsung dalam jangka panjang, mulai dari kehamilan hingga usia 24 bulan [1]. Balita dikategorikan stunting jika nilai *z-score* kurang dari -2 standar deviasi (pendek) dan kurang dari -3 standar deviasi (sangat pendek) [2]. Selama tahun 2022, sebanyak 148,1 juta (22,3%) anak di bawah usia 5 tahun di seluruh dunia mengalami stunting. Dari jumlah tersebut, 52% tinggal di Asia dan 43% di Afrika [3]. Berdasarkan data dari Kemendagri tahun 2023 menunjukkan bahwa di Indonesia, dari total 16,451,587 (92.9%) balita, sebanyak 1,172,051 (7.1%) di antaranya mengalami stunting. Prevalensi stunting menurut Survei Status Gizi Indonesia (SSGI) mengalami penurunan yaitu dari 24.4% pada tahun 2021 menjadi 21.6% pada tahun 2022. Pada tahun 2023 berdasarkan Survei Kesehatan Indonesia, angka stunting turun 0,1% menjadi 21.5%. Meskipun demikian, penurunan angka stunting di Indonesia belum mencapai standar WHO, yang menetapkan prevalensi stunting harus kurang dari 20%. Oleh karena itu, pemerintah Indonesia berharap penurunan prevalensi stunting dapat lebih signifikan, mencapai target 14% [4].

Menurunkan angka stunting sejak dini sangat penting untuk mencegah dampak negatif jangka panjang seperti gangguan perkembangan otak yang dapat mengurangi tingkat kecerdasan anak dan berpotensi mengurangi produktivitas di masa dewasa [5]. Oleh karena itu, perlu dilakukan prediksi balita stunting sebagai upaya preventif dalam menangani masalah tersebut. Salah satu metode prediksi adalah menggunakan teknik klasifikasi yang didasarkan pada beberapa atribut atau fitur terukur. Banyak metode klasifikasi pada *machine learning*, namun pada penelitian ini teknik klasifikasi untuk memprediksi balita stunting dilakukan dengan menggunakan *random forest*. Algoritma *random forest* adalah metode pemodelan berbasis pohon yang efektif dalam *machine learning*. Algoritma ini beroperasi dengan membangun beberapa pohon keputusan (*Decision Trees*) selama tahap penelitian. Setiap pohon dirancang menggunakan subset data dan fitur yang dipilih secara acak pada setiap proses pemisahan [6].

Dengan penggunaan jenis pengacakan yang sesuai, *random forest* mampu menjadi pengklasifikasi dan pemrediksi yang akurat [7]. Dalam prediksi, algoritma ini menggabungkan hasil dari semua pohon, baik dengan voting (untuk tugas klasifikasi) atau dengan averaging (untuk tugas regresi). Proses pengambilan keputusan kolaboratif ini menjadi keunggulan algoritma random forest yang akhirnya banyak digunakan untuk keperluan klasifikasi dan regresi. Selain itu random forest cukup populer karena kemampuannya menangani data kompleks, mengurangi overfitting, dan memberikan prakiraan yang andal berbagai keadaan.

Permasalahan muncul dalam menerapkan metode klasifikasi apapun adalah ketika data memiliki kelas tidak seimbang. Faktanya, sebagian besar dataset yang ditemukan dalam dunia nyata secara alami memiliki kelas yang tidak seimbang, misalnya dataset medis umumnya menghadapi masalah ketidakseimbangan ini [8]. Seperti yang telah disebutkan sebelumnya jumlah balita stunting jauh lebih sedikit dibandingkan dengan balita normal. Misalnya dalam 1000 balita hanya belasan balita yang mengalami stunting, ini artinya terjadi ketidakseimbangan ekstrem. Secara teoretis proses klasifikasi dengan algortima *machine learning* apapun akan sulit menghadapi data tidak seimbang ekstrem seperti ini karena tahapan learning melaui data training tidaklah cukup untuk data kelas minoritas, namun terdapat kemungkinan *over-fitting* pada data mayoritas. Selain itu, ketidakseimbangan kelas diketahui dapat meningkatkan bias terhadap kelas mayoritas [9].

Untuk itu perlu upaya untuk menjadikan kelas data seimbang. Menambah data alami tidaklah mungkin karena faktanya memang tidak tersedia. Untuk itu *resampling* adalah salah satu teknik yang sangat efektif dalam menangani masalah ketidakseimbangan kelas. Secara umum, teknik *resampling* terbagi menjadi tiga kategori, yaitu: (1) *oversampling*, yang meningkatkan jumlah sampel kelas minoritas dengan mereplikasi sampel baru dari kelas tersebut; (2) *undersampling*, yang mengurangi jumlah sampel dari kelas mayoritas untuk menyeimbangkan distribusi kelas dengan kelas minoritas; (3) *hybrid sampling*, yang menggabungkan kelebihan dari kedua teknik *oversampling* dan *undersampling* untuk mengatasi ketidakseimbangan kelas [10].

Penelitian [11] membangun Bayesian Networks yang dikembangkan oleh tiga algoritma (Tabu, Hill-climbing, dan MMHC) dengan seleksi fitur Boruta untuk data pemantauan Diabetes Mellitus dengan ketidakseimbangan kelas. Penelitian tersebut mengatasi ketidakseimbangan kelas menggunakan dua metode *oversampling* yaitu *Synthetic Minority Oversampling Technique* (SMOTE) dan *Borderline*-SMOTE, dan metode *hybrid sampling* yaitu *Synthetic Minority Oversampling Technique with Edited Nearest Neighbors* (SMOTE-ENN). Diperoleh hasil kombinasi Boruta-SMOTE-ENN-Tabu menunjukkan kinerja paling baik.

Pada penelitian ini akan akan dibandingkan dua metode *resampling*, yaitu SMOTE dan SMOTE-ENN sebelum algoritma random forest diterapkan untuk klasifikasi. SMOTE membangkitkan sampel sintetis dengan menginterpolasi *k-Nearest Neighbors* (KNN) dari masing-masing sampel minoritas [12]. SMOTE-ENN awalnya melakukan *oversampling* pada kelas minoritas menggunakan interpolasi dan kemudian menghapus sampel yang berlebihan menggunakan metode ENN. Akhirnya, algoritma ini menghasilkan data kelas yang seimbang [13]. Dalam penelitian ini simulasi klasifikasi dilakukan pada tiga skenario, yaitu data asli yang tidak seimbang, data seimbang hasil replikasi menggunakan SMOTE, dan data seimbang hasil replikasi menggunakan SMOTE-ENN. Selanjutnya, performa kinerja klasifikasi *random forest* dibandingkan untuk ketiga skenario tersebut menggunakan *confusion matrix* dan berbagai instrumen numerik lainnya.

2 METODE PENELITIAN

2.1 Pengumpulan Data

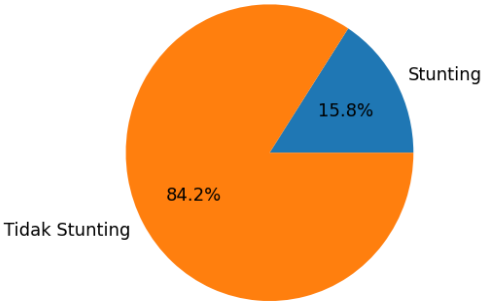
Data yang digunakan dalam penelitian ini adalah data hasil generate menggunakan pembangkit random, kemudian ditransformasi dalam *z-score*. Selanjutnya, penentuan atribut dan kelas label didasarkan pada peraturan Menkes Nomor 2 tahun 2020 tentang standar antropometri anak. Atribut yang dimaksud adalah umur (bulan), jenis kelamin, dan tinggi badan (cm). Data hasil generate yang dimaksud dapat diakses pada https://bit.ly/DataSintetis.

Sampel berjumlah 500 kasus di mana 421 (84,2%) balita tidak stunting dan sisanya 79 (15,8%) balita mengalami stunting. Tabel 1 menyajikan beberapa kasus klasifikasi stunting, sedangkan Gambar 1 menyajikan proporsi antara kelas mayoritas dan minoritas.

Table 1. Data Balita Stunting

	Umur (Bulan)	Jenis Kelamin	Tinggi Badan (cm)	Label Stunting
1	0	Perempuan	49	Tidak

2	0	Perempuan	51.7	Tidak
3	0	Perempuan	51.3	Tidak
⋮	⋮	⋮	⋮	⋮
11	1	Perempuan	48.3	Ya
⋮	⋮	⋮	⋮	⋮
499	60	Peremuan	109.7	Tidak
500	60	Laki-Laki	120.5	Tidak



Gambar 1. Proporsi Data Mayoritas dan Data Minoritas

2.2 Pra Proses Data

Pra proses data adalah tahap menyajikan data ke dalam format yang lebih mudah diproses menggunakan algoritma *machine learning*. Ada dua tahap pra proses data dalam penelitian ini, yaitu, pembersihan data untuk mendeteksi dan menghapus data yang tidak lengkap, salah entri, atau tidak relevan. Selanjutnya, data ditransformasi ke dalam bentuk numerik baik berupa scalar maupun vektor sehingga dapat diproses menggunakan paket (*library*) yang relevan pada Python [14].

2.3 Pembagian Data untuk Klasifikasi

Data dibagi menjadi dua kelompok yaitu data latih dan data uji. Model *machine learning* pada proses klasifikasi adalah *random forest* yang diterapkan pada data latih dan divalidasi menggunakan data uji. Rasio pembagian data latih dan data uji adalah 80:20. Artinya 80% data untuk data latih dan 20% untuk data uji. Model yang sudah dilatih digunakan untuk memprediksi kelas kasus baru di luar data latih dan data uji, yang disebut generalisasi model.

2.4 Random Forest

Random forest menggabungkan beberapa *decision tree* untuk meningkatkan akurasi dan sensitivitas terhadap data latih [15]. *Decision tree* adalah sebuah pengklasifikasi berstruktur pohon di mana atribut dataset disimpan di dalam simpul internal, aturan keputusan direpresentasikan oleh cabang-cabang, dan hasilnya dinyatakan oleh setiap simpul daun [16].

Terdapat tiga aspek penting dalam metode *random forest*, yakni: (1) penerapan *bootstrap sampling* untuk membangun pohon prediksi; (2) setiap pohon keputusan melakukan prediksi dengan menggunakan hasil dari masing-masing pohon keputusan melalui metode suara terbanyak (*voting*) untuk klasifikasi atau rata-rata (*averaging*) untuk regresi [17]. *Bootstrap sampling* adalah metode pengambilan sampel berbasis komputer yang melibatkan pemilihan objek dengan pengembalian, sehingga memungkinkan objek yang sama dipilih beberapa kali.

Dalam pemilihan atribut, *random forest* biasanya menggunakan *gini index* dan *entropy* sebagai kriteria pemilihan atribut [18]. Dalam penelitian ini digunakan *gini index* karena *gini index* lebih efisien secara komputasi dibandingkan dengan *entropy*. Misalkan S adalah sebuah himpunan yang terdiri dari s data. Data ini terbagi ke dalam m kelas yang berbeda (C_i, dengan $i = 1, \dots, m$). Berdasarkan kelas-kelas tersebut, S dapat dibagi menjadi m subset S_i dengan $i = 1, \dots, m$ di mana S_i adalah kumpulan data yang termasuk ke dalam kelas C_i, dan s_i adalah jumlah data dalam S_i [19]. Dengan demikian, *gini index* dapat dinyatakan dengan rumus berikut:

$$Gini\ index(S) = 1 - \sum_{i=1}^{m} \left(\frac{S_i}{S}\right)^2 \quad (1)$$

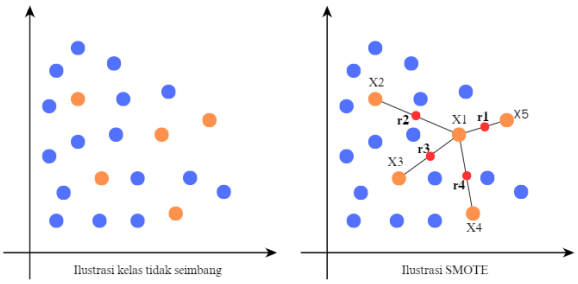
Gini index digunakan untuk memilih atribut yang akan ditetapkan sebagai *node* pada pohon. *Node* dipilih berdasarkan *gini index* terendah untuk mendapatkan pemisahan data terbaik. Pemisahan terus dilakukan hingga membentuk struktur pohon keputusan dan dapat digunakan untuk prediksi [20].

2.5 Imbalanced Class

Ketidakseimbangan kelas terjadi ketika ada lebih banyak sampel dari satu kelas (dikenal sebagai kelas mayoritas atau negatif) dalam kumpulan data daripada kelas lainnya (kelas minoritas atau positif). Secara umum, pendekatan klasifikasi standar mengasumsikan bahwa data latih memiliki distribusi yang seimbang. Oleh karena itu, dalam klasifikasi data yang tidak seimbang, metode ini cenderung didominasi oleh kelas mayoritas dan cenderung mengabaikan atau salah dalam mengklasifikasikan sampel dari kelas minoritas [21].

2.6 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE menggunakan teknik yang disebut interpolasi yaitu membuat titik data baru dalam rentang titik data yang telah diketahui. Kemudian menambah sampel kelas minoritas dengan menginterpolasi sampel sintetis. Hal ini mencegah duplikasi sampel minoritas dan menghasilkan sampel sintetis baru yang mirip dengan titik-titik yang diketahui [22]. Kelebihan utama dari SMOTE adalah tidak menyebabkan hilangnya informasi dan dapat meningkatkan akurasi prediksi untuk kelas minoritas [23].



Gambar 2. Ilustrasi SMOTE

Cara membangkitkan sampel sintetis dalam algoritma SMOTE diilustrasikan pada Gambar 2. Sebuah sampel positif x_1 dipilih sebagai titik dasar untuk membuat sampel sintetis yang baru. Sampel sintetis dibangkitkan berdasarkan *k-nearest neighbors* yang diperoleh dengan menghitung jarak *Euclid* antara data minoritas. Terakhir, interpolasi acak dilakukan untuk mendapatkan anggota sampel baru r_1 hingga r_4 [24]. Rumus untuk menghitung jarak *Euclid* menggunakan persamaan (2). Dan untuk membangkitkan sampel sintetis menggunakan persamaan (3).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

$$r = X_i + (\hat{X}_k - X_i) \times \delta, \quad (3)$$

di mana r : sampel sintetis, X_i : sampel ke-i dari kelas minoritas, $\hat{X}_k$: sampel dengan jarak terdekat dari X_i, dan δ : bilangan acak antara 0 dan 1. Jika δ mendekati 0 maka sampel sintetis akan mirip dengan data minoritas asal. Sebaliknya, jika δ mendekati 1 maka sampel sintetis akan mirip dengan data tetangga terdekat. Dan jika δ sekitar 0.5 maka ada kemungkinan sampel sintetis mirip dengan data mayoritas [25].

2.7 Synthetic Minority Oversampling Technique with Edited Nearest Neighbors (SMOTE-ENN)

Ide dasar SMOTE adalah membentuk sampel baru dari kelas minoritas dengan melakukan interpolasi di antara beberapa sampel kelas minoritas yang berdekatan. SMOTE dapat meningkatkan akurasi klasifikasi secara signifikan. Namun, SMOTE juga dapat menghasilkan *noise*. Algoritma SMOTE-ENN diusulkan Batista et. al. untuk meningkatkan kinerja SMOTE [26]. ENN digunakan untuk menghapus data dari kedua kelas. Setiap data yang label kelasnya berbeda setidaknya dua dari tiga tetangga terdekatnya akan dihapus dari data latih. Karena beberapa sampel kelas mayoritas mungkin memasuki ruang kelas minoritas dan sebaliknya. Sehingga kombinasi SMOTE dan ENN dapat mengurangi risiko *overfitting* yang disebabkan oleh data sintetis [27].

2.8 Performa Model

Untuk mengukur performa model digunakan *confusion matrix* yaitu metrik yang sering digunakan dalam pemecahan masalah klasifikasi, metrik ini dapat diterapkan pada masalah klasifikasi multi-kelas dan masalah klasifikasi biner. *Confusion matrix* merepresentasikan jumlah nilai yang diprediksi dan nilai aktual [28], seperti ditunjukkan pada Tabel 2.

Table 2. Confusion Matrix

Prediksi	**Aktual**	
	Positif (1)	Negatif (0)
Positif (1)	TP	FP
Negatif (0)	FN	TN

Performa model dihitung berdasarkan nilai akurasi, presisi, recall, dan F_1-*score*.

1. Akurasi adalah rasio antara prediksi yang benar (positif maupun negatif dengan total data yang ada.

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

2. Presisi adalah proporsi prediksi positif yang benar dibandingkan dengan seluruh hasil prediksi positif.

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

3. Recall adalah proporsi prediksi positif yang benar dibandingkan dengan total data yang sebenarnya positif.

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

4. F_1-*score* adalah rata-rata harmonic antara presisi dan recall, yang mempertimbangkan bobot keduanya.

$$F_1 - Score = 2 \times \frac{Recall \times Presisi}{Recall + Presisi} \quad (7)$$

3 HASIL DAN PEMBAHASAN

Data yang akan digunakan dalam penelitian ini masih berupa data mentah sehingga perlu dilakukan tahap pra proses data. Pada penelitian ini akan dilakukan 2 tahap pra proses data yaitu transformasi data dan pembersihan data dari *missing value* dan data duplikat.

Transformasi data yaitu mengubah data kategorik menjadi data numerik, seperti “Perempuan”, “Laki-laki” menjadi “0” dan “1”, dan pada variabel Stunting seperti “Tidak”, “Ya”, diubah menjadi “0” dan “1”. Selanjutnya, tahap pembersihan data. Dalam data tersebut tidak ditemukan *missing value*, namun terdapat 6 data duplikat. Sehingga data yang akan digunakan untuk pemodelan sebanyak 494 baris data.

Table 3. Hasil Pra Proses Data

	Umur (Bulan)	Jenis Kelamin	Tinggi Badan (cm)	Stunting
1	0	0	49	0
2	0	0	51.7	0
3	0	0	51.3	0
⋮	⋮	⋮	⋮	⋮
11	1	0	48.3	1
⋮	⋮	⋮	⋮	⋮
499	60	0	109.7	0
500	60	1	120.5	0

Tahap selanjutnya adalah membagi data menjadi dua dengan perbandingan 80:20, diperoleh data latih sebanyak 395 dan data uji sebanyak 99. Mengingat adanya ketidakseimbangan kelas dalam data balita stunting, maka metode SMOTE dan SMOTE-ENN akan diterapkan pada data latih untuk menangani masalah tersebut menggunakan *library imblearn.over_sampling* untuk SMOTE dan *imblearn.combine* untuk SMOTE-ENN. Sehingga penelitian ini akan melakukan tiga skenario, yaitu: (1) Klasifikasi menggunakan *random forest* tanpa mempertimbangkan ketidakseimbangan kelas; (2) SMOTE digunakan untuk menambah kelas minoritas; (3) SMOTE-ENN digunakan untuk menambah kelas minoritas dan menghapus data yang tidak sesuai.

Table 4. Hasil Penerapan SMOTE dan SMOTE-ENN

Stunting	Data Awal	Data dengan SMOTE	Data dengan SMOTE-ENN
0 = Tidak	329	329	318
1 = Ya	66	329	311

Tabel 4. Menunjukkan bahwa SMOTE dapat membentuk data sintetis untuk kelas minoritas “Ya” (*Stunting*) sebanyak 263. Sedangkan, SMOTE-ENN dapat membentuk data sintetis sebanyak 245 dan menghapus data yang tidak sesuai pada kelas mayoritas “Tidak” (*Tidak stunting*) sebanyak 11.

Table 5. Performa Model

	Akurasi	Presisi	Recall	F1-Score
Tanpa *Resampling Data*	89%	78%	64%	68%
SMOTE	91%	80%	78%	79%
SMOTE-ENN	90%	79%	71%	74%

Pada Tabel 5. Menunjukkan bahwa *random forest* efektif dalam memprediksi balita stunting dengan nilai akurasi sebesar 89%. Selain itu, penerapan SMOTE dan SMOTE-ENN mampu meningkatkan kinerja klasifikasi dengan nilai akurasi masing-masing sebesar 91% dan 90%. Meskipun SMOTE-ENN dapat memperbaiki masalah *overfitting* dengan menghapus sampel yang dianggap *noise* melalui ENN, ini juga dapat menyebabkan hilangnya data yang relevan. Di sisi lain, SMOTE lebih stabil dalam mempertahankan variasi data dengan hanya menambahkan sampel tanpa menghapus data asli. Sehingga berdasarkan nilai presisi, *recall*, dan F_1*-score* terlihat bahwa SMOTE memberikan hasil yang lebih baik dibandingkan SMOTE-ENN.

Table 6. Confusion Matrix Random Forest

Aktual	Prediksi	
	0	1
0	84	2
1	9	4

Berdasarkan Tabel 6, *random forest* dapat memprediksi balita normal dengan benar sebanyak 84, dan balita stunting sebanyak 4. Namun, model salah memprediksi 2 balita normal sebagai balita stunting, dan 9 balita stunting sebagai balita normal.Sehingga diperoleh nilai presisi 78%, nilai *recall* 64%, dan nilai F_1-*score* 68%.

Table 7. Confusion Matrix Random Forest+SMOTE

Aktual	Prediksi	
	0	1
0	82	4
1	5	8

Berdasarkan Tabel 7, *random forest* + SMOTE dapat memprediksi balita normal dengan benar sebanyak 82, dan balita stunting sebanyak 8. Namun, model salah memprediksi 4 balita normal sebagai balita stunting, dan 5 balita stunting sebagai balita normal.Sehingga diperoleh nilai presisi 80%, nilai *recall* 78%, dan nilai F_1-*score* 79%.

Table 8. Confusion Matrix Random Forest+SMOTE-ENN

Aktual	Prediksi	
	0	1
0	83	3
1	7	6

Berdasarkan Tabel 8, *random forest* + SMOTE-ENN dapat memprediksi balita normal dengan benar sebanyak 83, dan balita stunting sebanyak 6. Namun, model salah memprediksi 3 balita normal sebagai balita stunting, dan 7 balita stunting sebagai balita normal.Sehingga diperoleh nilai presisi 79%, nilai *recall* 71%, dan nilai F_1-*score* 74%.

4 KESIMPULAN

Berdasarkan simulasi tiga skenario klasifikasi diperoleh hasil bahwa SMOTE paling baik disusul SMOTE-ENN dan klasifikasi tanpa resampling data. Klasifikasi *random forest* menghasilkan nilai akurasi sebesar 89%, nilai presisi sebesar 78%, nilai *recall* sebesar 64%, dan nilai *f1-score* sebesar 68%. SMOTE meningkatkan nilai akurasi sebesar 2%, nilai presisi sebesar 2%, nilai *recall* sebesar 14%, dan nilai *f1-score* sebesar 11%. SMOTE-ENN meningkatkan akurasi sebesar 1%, nilai presisi sebesar 1%, nilai *recall* sebesar 7%, dan nilai *f1-score* sebesar 6%. Untuk meningkatkan performa klasifikasi *random forest* dengan SMOTE disarankan menambahkan atribut sebagai kriteria stunting.

5 KONTRIBUSI PENELITIAN

Hasil penelitian ini diharapkan dapat memberikan kontribusi positif terhadap pengembangan sistem deteksi dini stunting, yang sangat penting untuk memungkinkan intervensi Kesehatan yang lebih cepat dan tepat. Selain itu, penelitian ini turut memperkaya literatur di bidang *machine learning* dengan penerapan langsung pada dataset kesehatan yang relevan, serta menyediakan solusi yang dapat diterapkan untuk mengatasi masalah klasifikasi data tak seimbang di berbagai bidang lainnya.

DAFTAR PUSTAKA

[1] R. Hitman *et al.*, "Penyuluhan Pencegahan Stunting pada Anak (Stunting Prevention Expansion in Children)," *Communnity Development Journal*, vol. 2, no. 3, 2021.

[2] E. Lestari, Z. Shaluhiyah, and M. Sakundarno Adi, "MPPKI Media Publikasi Promosi Kesehatan Indonesia," vol. 6, no. 2, 2023, doi: 10.31934/mppki.v2i3.

[3] Unicef, WHO, and World Bank, *level and trends in child malnutrition 2023*. 2023.

[4] Rokom, "Prevalensi Stunting di Indonesia Turun ke 21,6% dari 24.4%." Accessed: May 17, 2024. [Online]. Available: https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20230125/3142280/prevalensi-stunting-di-indonesia-turun-ke-216-dari-244/

[5] "Pedoman Pelaksanaan Intervensi Penurunan Stunting Terintegrasi Di Kabupaten Kota".

[6] "Random Forest Algorithm in Machine Learning." Accessed: Dec. 09, 2024. [Online]. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning

[7] L. Breiman, "Random Forests," 2001.

[8] V. Kumar *et al.*, "Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques," *Healthcare (Switzerland)*, vol. 10, no. 7, Jul. 2022, doi: 10.3390/healthcare10071293.

[9] T. Bouabana-Tebibel and S. H. Rubin, "Advances in Intelligent Systems and Computing 446." [Online]. Available: http://www.springer.com/series/11156

[10] R. Ghorbani and R. Ghousi, "Comparing Different Resampling Methods in Predicting Students' Performance Using Machine Learning Techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

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

[12] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," 2002.

[13] M. Muntasir Nishat *et al.*, "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset," *Sci Program*, vol. 2022, 2022, doi: 10.1155/2022/3649406.

[14] D. Varma, A. Nehansh, and P. Swathy, "Data Preprocessing Toolkit : An Approach to Automate Data Preprocessing," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 07, no. 03, Mar. 2023, doi: 10.55041/ijsrem18270.

[15] S. Das, M. S. Imtiaz, N. H. Neom, N. Siddique, and H. Wang, "A hybrid approach for Bangla sign language recognition using deep transfer learning model with random forest classifier," *Expert Syst Appl*, vol. 213, Mar. 2023, doi: 10.1016/j.eswa.2022.118914.

[16] G. Devisetty and N. S. Kumar, "Prediction of Bradycardia using Decision Tree Algorithm and Comparing the Accuracy with Support Vector Machine," in *E3S Web of Conferences*, EDP Sciences, Jul. 2023. doi: 10.1051/e3sconf/202339909004.

[17] A. Primajaya and B. N. Sari, "Random Forest Algorithm for Prediction of Precipitation," 2018.

[18] C. Zhang, Y. Liu, and N. Tie, "Forest Land Resource Information Acquisition with Sentinel-2 Image Utilizing Support Vector Machine, K-Nearest Neighbor, Random Forest, Decision Trees and Multi-Layer Perceptron," *Forests*, vol. 14, no. 2, Feb. 2023, doi: 10.3390/f14020254.

[19] T. Setiyorini *et al.*, "Penerapan Gini Index dan K-Nearest Neighbor untuk Klasifikasi Tingkat Kognitif Soal pada Taksonomi Bloom," *Jurnal Pilar Nusa Mandiri*, vol. 13, no. 2, 2017, [Online]. Available: http://www.nusamandiri.ac.id1;http://www.swadharma.ac.id/2

[20] L. C, P. S, A. H. Kashyap, A. Rahaman, S. Niranjan, and V. Niranjan, "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers," *Cancer Inform*, vol. 22, Jan. 2023, doi: 10.1177/11769351231167992.

[21] P. Soltanzadeh and M. Hashemzadeh, "RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Inf Sci (N Y)*, vol. 542, pp. 92–111, Jan. 2021, doi: 10.1016/j.ins.2020.07.014.

[22] K. Abhishek and M. Abdelaziz, *Machine learning for imbalanced data : tackle imbalanced datasets using machine learning and deep learning techniques*.

[23] N. P. Y. T. Wijayanti, E. N. Kencana, and I. W. Sumarjaya, "SMOTE: Potensi dan Kekurangannya pada Survei," *E-Jurnal Matematika*, vol. 10, no. 4, p. 235, Nov. 2021, doi: 10.24843/mtk.2021.v10.i04.p348.

[24] A. Salvadorrgarcíaa, M. R. Pratii, and B. Franciscooherrera, "Learning from Imbalanced Data Sets."

[25] B. Santoso, H. Wijayanto, K. A. Notodiputro, and B. Sartono, "Synthetic over Sampling Methods for Handling Class Imbalanced Problems : A Review," in *IOP Conference Series: Earth and Environmental Science*, Institute of Physics Publishing, Apr. 2017. doi: 10.1088/1755-1315/58/1/012031.

[26] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data."

[27] K. Wang *et al.*, "Improving risk identification of adverse outcomes in chronic heart failure using smote +enn and machine learning," *Risk Manag Healthc Policy*, vol. 14, pp. 2453–2463, 2021, doi: 10.2147/RMHP.S310295.

[28] A. Kulkarni, F. A. Batarseh, and D. Chong, "Chapter 5: Foundations of Data Imbalance and Solutions for a Data Democracy."