FROM DATA IMBALANCE TO PRECISION: SMOTE-DRIVEN MACHINE LEARNING FOR EARLY DETECTION OF KIDNEY DISEASE

OPTIMASI KLASIFIKASI DATA TIDAK SEIMBANG PADA DATASET MEDIS PADA KASUS PENYAKIT GINJAL KRONIS DENGAN TEKNIK SMOTE

R.M Aldani Adi Bhirawa[1*], Ucta Pradema Sanjaya[2],
[1,2]Informatics Engineering Study Programme, Faculty of Computer Science and Education, Ngudi Waluyo University, Ungaran, Central Java, Indonesia
E-mail: [1]*rmaldaniadibhirawa@gmail.com*, [2]*uctapradema@unw.ac.id*

Abstract- *Chronic Kidney Disease (CKD) has become a significant global health issue, with its prevalence rising sharply, particularly in developing countries like Indonesia. According to the Kementrian Kesehatan (KEMENKES), the Synthetic Minority Over-sampling Technique (SMOTE) has been widely adopted to address this. SMOTE generates synthetic samples for the minority class, enhancing the model's ability to identify high-risk patients. Studies demonstrate SMOTE's effectiveness, particularly when combined with ensemble learning algorithms like Random Forest and Gradient Boosting. The data collection focused on relevant medical parameters critical for the study, encompassing laboratory test results, diagnostic reports, and clinical observations related to kidney function. This dataset in kidney disease is used to predict whether someone has chronic kidney disease or not with a total sample of 400 data obtained from the Ungaran Regional Hospital and several clinics that can detect kidney disease. Recent research highlights that SMOTE significantly improves model accuracy, with Random Forest achieving 99.30% accuracy. These findings emphasise the importance of data balancing in enhancing diagnostic precision, offering promising avenues for early CKD detection and improved patient outcomes.*

Keywords: *Gradient Boosting, Chronic Kidney, SMOTE, Random Forest, Gradient Boosting.*

Abstrak- Penyakit Ginjal Kronis (PGK) telah menjadi masalah kesehatan global yang signifikan, dengan prevalensi yang meningkat tajam, terutama di negara-negara berkembang seperti Indonesia. Menurut Kementrian Kesehatan (KEMENKES), untuk mengatasi hal ini, Teknik Pengambilan Sampel Over-sampling Minoritas Sintetis (SMOTE) telah diadopsi secara luas. SMOTE menghasilkan sampel sintetis untuk kelas minoritas, sehingga meningkatkan kemampuan model untuk mengidentifikasi pasien berisiko tinggi. Penelitian menunjukkan efektivitas SMOTE, terutama ketika dikombinasikan dengan algoritme pembelajaran ensembel seperti Random Forest dan Gradient Boosting. Pengumpulan data difokuskan pada parameter medis yang relevan dan penting untuk penelitian, yang meliputi hasil tes laboratorium, laporan diagnostik, dan pengamatan klinis yang berkaitan dengan fungsi ginjal. Dataset penyakit ginjal ini digunakan untuk memprediksi apakah seseorang menderita penyakit ginjal kronis atau tidak dengan total sampel sebanyak 400 data yang diperoleh dari rumah sakit umum daerah ungaran dan beberapa klinik yang dapat mendeteksi penyakit ginjal. Penelitian terbaru menyoroti bahwa SMOTE secara signifikan meningkatkan akurasi model, dengan Random Forest mencapai akurasi 99,30%. Temuan ini menekankan pentingnya penyeimbangan data dalam meningkatkan ketepatan diagnostik, menawarkan jalan yang menjanjikan untuk deteksi dini CKD dan meningkatkan hasil pasien.

Keyword: Gradient Boosting, Penyakit Ginjal Kronis, SMOTE, Random Forest, Gradient Boosting.

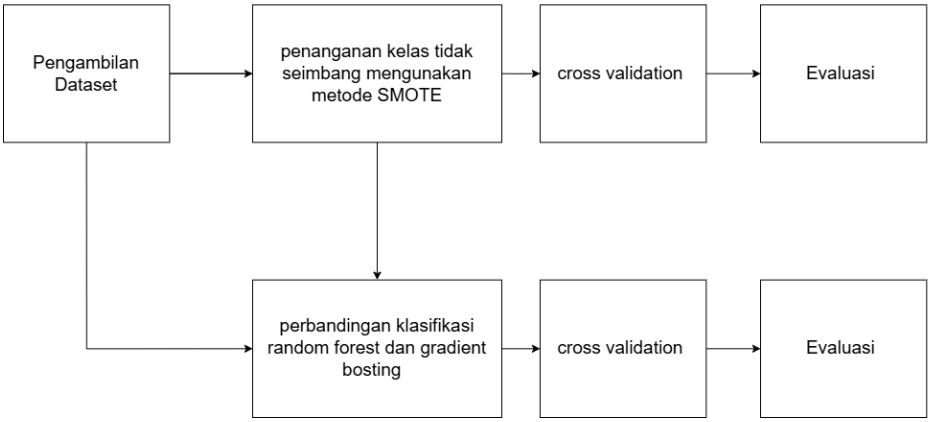
I. PENDAHULUAN

Upaya untuk mengurangi jumlah penderita penyakit ginjal jinak bertujuan untuk memberikan pengobatan dini yang cepat dan tepat guna mencegah perkembangan penyakit ginjal menjadi lebih serius. Secara umum, banyak orang tidak mengetahui bahwa ginjal memainkan peran besar dalam mengatur tubuh kita[1]. Ginjal berfungsi untuk memproduksi urin dan menjaga keseimbangan cairan; misalnya, dalam suhu dingin, tubuh cenderung lebih sering buang air kecil, sedangkan dalam kondisi panas, tubuh mengalami kekurangan cairan yang lebih besar[2]. Dr. Zulkhair Ali, Ketua Pernefri (Perhimpunan Nefrologi Indonesia), menekankan bahwa disfungsi ginjal dapat menyebabkan gagal ginjal[3]. Kurangnya informasi kepada masyarakat mengenai gejala penyakit ginjal turut memperburuk masalah ini. Oleh karena itu, sangat penting untuk menerapkan langkah-langkah pencegahan dengan mengidentifikasi individu yang berisiko mengalami penyakit ginjal[3]–[5]. Pencegahan idealnya dimulai sejak tahap awal, di mana pasien menjalani skrining untuk mengetahui ada atau tidaknya faktor risiko penyakit ginjal. Dengan kemajuan teknologi, inovasi telah merambah ke berbagai sektor, termasuk industri kesehatan. Dalam institusi kesehatan, ketepatan analisis dan efisiensi waktu sangat penting untuk pengambilan keputusan yang tepat. Dalam dunia medis, klasifikasi penyakit berbasis kecerdasan buatan (AI) telah berkembang pesat untuk meningkatkan akurasi diagnosis dan membantu tenaga medis dalam pengambilan keputusan[2]–[9]. Namun, salah satu tantangan utama dalam penerapan model klasifikasi adalah ketidakseimbangan data dalam dataset medis[2], [10]–[15]. Data pasien yang positif mengidap penyakit ginjal kronis (PGK) sering kali jauh lebih sedikit dibandingkan dengan data pasien sehat, sehingga model machine learning cenderung lebih bias terhadap kelas mayoritas. Ketidakseimbangan ini dapat mengurangi kemampuan model dalam mengidentifikasi pasien berisiko tinggi secara akurat[8], [16], [17].

Berbagai metode telah dikembangkan untuk menangani masalah ketidakseimbangan data ini. Salah satu metode yang paling banyak digunakan adalah Synthetic Minority Over-sampling Technique (SMOTE)[18]–[20]. Teknik ini bekerja dengan menghasilkan sampel sintetis berdasarkan kedekatan data minoritas untuk meningkatkan representasi kelas tersebut dalam dataset. Dengan menggunakan SMOTE, diharapkan kinerja model klasifikasi dapat meningkat dalam mendeteksi pasien yang sebenarnya menderita Penyakit Ginjal Kronis (PGK). Beberapa penelitian sebelumnya telah membuktikan efektivitas SMOTE dalam menangani dataset yang tidak seimbang. Sebagai contoh, studi oleh Garcia et al. menunjukkan bahwa penerapan SMOTE pada dataset medis dapat meningkatkan nilai recall tanpa mengorbankan akurasi secara signifikan[21]. Studi lain menemukan bahwa kombinasi SMOTE dengan algoritma ensemble learning, seperti Random Forest dan Gradient Boosting, dapat memberikan hasil yang lebih akurat dibandingkan dengan metode oversampling lainnya.

Selain itu, penelitian menunjukkan bahwa meskipun SMOTE efektif dalam meningkatkan kinerja klasifikasi, pemilihan parameter yang tepat dalam teknik ini sangat memengaruhi hasil akhir[11], [20]. Jika tidak disesuaikan dengan baik, SMOTE dapat menyebabkan overfitting pada model, sehingga mengurangi kemampuan generalisasi terhadap data baru. Oleh karena itu, diperlukan penelitian lebih lanjut untuk mengeksplorasi implementasi SMOTE yang optimal dalam kasus CKD, guna menghasilkan model klasifikasi yang lebih andal dan akurat. Mengingat urgensi deteksi dini CKD dan tantangan dalam klasifikasi data medis yang tidak seimbang, penelitian ini berfokus pada penerapan SMOTE dalam optimalisasi model klasifikasi PGK. Melalui penelitian ini, diharapkan dapat diperoleh pemahaman yang lebih baik tentang dampak ketidakseimbangan data terhadap model klasifikasi, serta efektivitas SMOTE dalam meningkatkan akurasi diagnosis CKD secara otomatis.

II. METODOLOGI PENELITIAN



Gambar 1 Alur penelitian

Alur penelitian ini mengimplementasikan strategi komputasi untuk mengatasi class imbalance melalui sintesis data dan validasi ketat. Dataset awal diambil dari repositori publik atau hasil akuisisi eksperimen, menjalani missing value handling dan normalisasi menggunakan teknik seperti Min-Max sebelum diproses. SMOTE diterapkan secara selektif pada minority class dengan menghasilkan sampel sintetis melalui interpolasi linier antara instances terdekat dalam feature space, mencegah overfitting artifisial yang sering muncul pada oversampling konvensional. Proses validasi mengadopsi stratified k-fold cross-validation dengan k=5 untuk mempertahankan distribusi kelas asli selama pembagian fold, memastikan evaluasi model merepresentasikan kinerja pada skenario data nyata. Setiap iterasi melatih Random Forest dengan entropy-based splitting dan Gradient Boosting menggunakan learning rate adaptive, kemudian melakukan blind testing pada hold-out fold. Metrik evaluasi difokuskan pada precision-recall tradeoff dan F1-score sebagai primary metric mengingat kelemahan akurasi dalam konteks imbalanced data, dilengkapi analisis AUC-ROC untuk mengukur class separability.

A. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini berisi informasi medis yang terkait dengan Penyakit Ginjal Kronis (CKD). Dataset ini terdiri dari 400 rekam medis pasien dengan 24 karakteristik yang mencakup parameter klinis seperti kadar hemoglobin, tekanan darah, dan kadar serum kreatinin. Distribusi kelas dalam dataset ini tidak seimbang, di mana jumlah pasien yang menderita CKD lebih sedikit dibandingkan dengan pasien yang tidak menderita penyakit ini. Untuk mengatasi ketidakseimbangan ini, digunakan metode Synthetic Minority Oversampling Technique (SMOTE). Dataset ini digunakan untuk memprediksi apakah seseorang menderita penyakit ginjal kronis atau tidak, dengan total 400 sampel yang diperoleh dari Rumah Sakit Daerah Ungaran dan beberapa klinik yang dapat mendeteksi penyakit ginjal [3], [14], [15]. Data yang dikumpulkan mencakup beberapa parameter sebagai berikut: Usia (age), Tekanan Darah (bp), Gravitasi Spesifik (sg), Albumin (al), Gula Darah (su), Sel Darah Merah (rbc), Kadar Nitrogen Urea Darah (bun), Serum Kreatinin (sc), Sodium (sod), Kalium (pot), Hemoglobin (hemo), Tekanan Darah Tinggi (htn), Diabetes Mellitus (dm), Penyakit Arteri Koroner (cad), Anemia (ane), Edema (pe), Kelelahan (pcv), Nafsu Makan Buruk (appet), Nyeri Punggung (ba), Kram Otot (musa), Bau Urine Tidak Normal (wc), Hipertensi (ht). Dan alasan utama ketidakseimbangan dalam dataset ini adalah karena tidak semua orang menjalani pemeriksaan ginjal, sehingga data pasien dengan CKD jauh lebih sedikit dibandingkan pasien yang tidak menderita penyakit ini.

B. Kelas Tidak Seimbang

Data yang tidak seimbang mengacu pada situasi di mana distribusi kelas dalam dataset tidak merata, dengan beberapa kelas memiliki jumlah sampel yang jauh lebih sedikit atau lebih banyak dibandingkan kelas lainnya [8], [17]. Kelompok data dengan jumlah sampel lebih sedikit disebut sebagai kelompok minoritas, sedangkan kelompok dengan jumlah sampel lebih banyak disebut sebagai kelompok mayoritas. Akurasi sering digunakan sebagai metrik utama dalam mengevaluasi performa klasifikasi. Namun, dalam konteks ketidakseimbangan kelas, metrik ini kurang cocok, karena kelas minoritas hanya memberikan kontribusi kecil terhadap total akurasi [22], [23].

C. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE digunakan untuk mengatasi noise kelas dalam rekam data CKD. Teknik ini menciptakan sampel sintetis baru berdasarkan interpolasi data minoritas yang sudah ada[24]. Dengan cara ini, SMOTE menghindari masalah overfitting yang sering terjadi pada metode oversampling sederhana. SMOTE merupakan teknologi oversampling yang dirancang untuk mengatasi masalah ketidakseimbangan kelas.Cara kerjanya adalah membuat sampel sintetis baru untuk kelas minoritas, yaitu kelas yang memiliki jumlah sampel lebih sedikit dibandingkan kelas mayoritas[25], [26].

D. Classification

Random Forest (RF) adalah metode klasifikasi dalam statistik komputasi. Metode klasifikasi digunakan untuk menganalisis berbagai fitur yang mengklasifikasikan setiap data ke dalam kategori kelas yang telah ditentukan. Seiring dengan perkembangan era big data, penggunaan metode statistik berbasis komputasi menjadi semakin luas. Metode RF dipilih karena memiliki kesalahan klasifikasi lebih kecil, akurasi klasifikasi lebih tinggi, mampu menangani data dalam jumlah besar, dan efektif dalam mengatasi data yang tidak lengkap. Sebagai teknik klasifikasi yang handal, Random Forest dapat menangani banyak variabel input tanpa mengalami masalah overfitting dengan mudah [18], [27]. Keunggulannya terletak pada kemampuannya mengurangi korelasi antar keputusan, sehingga menghasilkan prediksi yang lebih konsisten. Selain itu, fleksibilitasnya memungkinkan algoritma ini digunakan dalam berbagai tugas, baik untuk klasifikasi maupun regresi, serta mengidentifikasi fitur penting dalam dataset pelatihan. Model yang dihasilkan dievaluasi menggunakan confusion matrix [27].

Berikut adalah rumus random forest classification

$$Gini(t) = t\ 1 - \sum_{i=1}^{c}(pi)^2 \qquad (1)$$

Keunggulan Random Forest mengurangi overfitting karena menggunakan bagging dan pemilihan fitur secara acak. Dapat menangani data dengan fitur numerik maupun kategorikal. Robust terhadap outlier dan noise dalam data. Menyediakan estimasi kepentingan fitur [28], [29] dan Gradient Boosting adalah algoritma machine learning yang dikategorikan dalam ensemble learning, yang memanfaatkan metodologi boosting. Keduanya bekerja dengan membangun model secara berurutan, di mana setiap model berikutnya berusaha memperbaiki kesalahan (residual) dari pendahulunya. XGBoost merupakan iterasi yang lebih unggul dan dioptimalkan dari Gradient Boosting, dengan penambahan regularisasi, peningkatan komputasi, serta fitur tambahan untuk meningkatkan kecepatan dan akurasi [27], [28].

Gradient Boosting bekerja dengan membangun model secara bertahap. Setiap model berikutnya memperoleh residual (kesalahan prediksi) dari model sebelumnya. Proses ini dilakukan dengan meminimalkan fungsi loss melalui gradient descent [30]

1. Inisialisasi: Mulai dengan model awal, biasanya berupa prediksi konstan:

$$ff_0(x) = arg\ min_y \sum_{i-1}^{n} L(yi, y) \qquad (2)$$

di mana L adalah fungsi loss (misalnya, Mean Squared Error untuk regresi atau Log Loss untuk klasifikasi).

2. Iterasi Boosting:
   For each iteration m=1,2,…,Mm=1,2,…,M:
   Hitung residual (kesalahan prediksi) dari model sebelumnya:

$$r = -\left[\frac{aL\ (yi, Fm-1(xi))}{aFm-1(xi)}\right] \qquad (3)$$

   Latih model baru hm(x) untuk prediksi residual *r im*.
   Update model

$$f_m(x) = f_{m-1}(x) + v.H_m(x) \qquad (4)$$

3. Output final
   Setelah M iterasi, final model:

$$f_m(x) = f_o(x) + v \sum_{m=1}^{m} hm^{(x)} \qquad (5)$$

E. Evaluasi Model

1) Cross-validation

Cross-validation adalah teknik yang digunakan untuk mengevaluasi kinerja model dengan membagi dataset menjadi beberapa subset (folds) guna mengurangi bias dalam pengukuran akurasi. Teknik yang umum digunakan adalah k-fold cross-validation, di mana dataset dibagi menjadi k bagian yang seimbang. Pada setiap iterasi, satu bagian digunakan sebagai data uji, sementara sisanya digunakan untuk melatih model. Proses ini diulang sebanyak k kali, sehingga setiap bagian berkesempatan menjadi data uji satu kali. Hasil akhirnya merupakan rata-rata dari semua iterasi, sehingga memberikan estimasi yang lebih stabil dan akurat terhadap performa model[26], [31].

Rumus perhitungan akurasi dengan k-fold cross-validation:

$$Accuracy_{CV} = \frac{1}{k} \sum_{i=1}^{k} Accuracy_i \qquad (6)$$

di mana:

- k adalah jumlah fold dalam cross-validation
- $Accuracy_i$ adalah akurasi yang diperoleh dari fold ke-i

Semakin besar nilai k, semakin kecil kemungkinan model mengalami bias, tetapi juga meningkatkan waktu komputasi. Umumnya, k dipilih dalam rentang 5 hingga 10. Teknik ini digunakan untuk memastikan bahwa model tidak hanya bekerja baik pada data pelatihan tetapi juga memiliki performa yang konsisten pada data yang tidak terlihat sebelumnya.

Dalam konteks klasifikasi penyakit ginjal kronis, cross-validation membantu menghindari overfitting dan memastikan bahwa model yang dibangun memiliki generalisasi yang baik terhadap data baru. Teknik ini sangat bermanfaat terutama pada dataset yang tidak seimbang, seperti yang telah diatasi dengan SMOTE.

2) Evaluasi matrix

Matrix dan laporan klasifikasi. Metrik-metrik ini dipilih karena kemampuannya dalam memberikan analisis komprehensif terhadap kinerja model pada berbagai kelas, yang sangat penting untuk keakuratan diagnosis medis. Confusion matrix adalah tabel yang menggambarkan kinerja model klasifikasi pada dataset uji dengan nilai sebenarnya yang sudah diketahui[32]. Tabel ini

memfasilitasi visualisasi prediksi model, menunjukkan prediksi yang benar dan salah untuk setiap kelas. Matriks ini disusun sedemikian rupa sehingga setiap baris mewakili instance dari kelas aktual, sementara setiap kolom menunjukkan instance dari kelas yang diprediksi :

- True Positive (TP): Prediksi positif yang benar.
- False Positive (FP): Salah diklasifikasikan sebagai positif.
- True Negative (TN): Prediksi negatif yang benar.
- False Negative (FN): Salah diklasifikasikan sebagai negatif.

Dari nilai-nilai ini, berbagai metrik kinerja seperti akurasi, presisi, recall, dan F1-score dapat diekstrak dalam laporan klasifikasi, sebagaimana diilustrasikan dalam Persamaan 6-9.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+fP} \quad (7)$$

$$\text{Precision} = \frac{TP}{FP+TP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (9)$$

$$\text{F1-Score } 2X \ \frac{Precision \times Recall}{Precision+Recall} \quad (10)$$

III. HASIL DAN PEMBAHASAN

Hasil penelitian ini memberikan wawasan yang signifikan dalam analisis dan klasifikasi penyakit ginjal, khususnya dalam deteksi dini serta tantangan yang terkait dengan dataset medis yang tidak seimbang. Bagian ini menguraikan temuan yang diperoleh dari data yang dikumpulkan di RSUD Kabupaten Semarang dan berbagai klinik di wilayah tersebut, diikuti dengan diskusi mendalam mengenai implikasinya. Analisis ini meneliti efektivitas metode yang digunakan, termasuk penerapan SMOTE untuk mengatasi ketidakseimbangan data, serta menilai pengaruhnya dalam meningkatkan akurasi klasifikasi penyakit ginjal. Diskusi ini bertujuan untuk memperkaya pengembangan instrumen diagnostik yang lebih andal serta strategi pengelolaan penyakit ginjal di lingkungan kesehatan.

A. Teknik Pengumpulan Data

Data dalam penelitian ini dikumpulkan di Rumah Sakit Umum Daerah (RSUD) Kabupaten Semarang serta beberapa klinik di sekitarnya. Data terutama bersumber dari catatan laboratorium dan laporan klinis, dengan tetap mematuhi peraturan etika dan kesehatan yang berlaku di Indonesia. Proses pengumpulan data dilakukan dengan ketat mengikuti prinsip kerahasiaan dan perlindungan data, sehingga tidak ada informasi pasien yang bersifat pribadi atau dapat diidentifikasi dalam dataset.

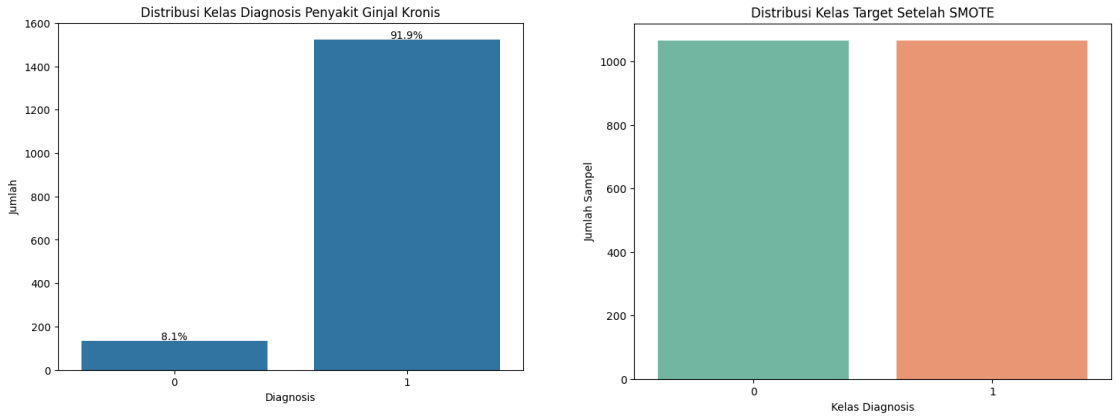
Pendekatan ini memastikan kepatuhan terhadap standar etika penelitian medis, dengan mengutamakan kerahasiaan pasien dan perlindungan data. Pengumpulan data difokuskan pada parameter medis yang relevan dengan penelitian ini, mencakup hasil tes laboratorium, laporan diagnostik, serta observasi klinis terkait fungsi ginjal. Dataset juga dianonimkan untuk menghapus informasi pribadi, sehingga sesuai dengan regulasi kesehatan Indonesia serta standar etika penelitian.

Kolaborasi dengan RSUD Kabupaten Semarang dan klinik lokal memungkinkan pengembangan dataset yang komprehensif dan mencerminkan kondisi pasien secara akurat di wilayah tersebut. Pendekatan ini meningkatkan validitas penelitian dan memastikan bahwa hasil yang diperoleh relevan dengan konteks layanan kesehatan setempat. Kepatuhan terhadap prinsip etika serta regulasi kesehatan menegaskan komitmen terhadap praktik penelitian yang bertanggung jawab, yang merupakan aspek krusial dalam studi medis.

B. Penanganan dengan Metode SMOTE

Kelas diagnosis penyakit ginjal memiliki distribusi yang tidak seimbang, dengan persentase diagnosis positif sebesar 9,1%. Ini berarti bahwa hanya sekitar 9,1% dari total data yang termasuk dalam kategori positif (menderita penyakit ginjal kronis), sementara 90,9% sisanya termasuk dalam kategori negatif (tidak menderita penyakit ginjal kronis).

Ketidakseimbangan ini menunjukkan bahwa dataset yang digunakan bersifat imbalanced, di mana salah satu kelas (negatif) mendominasi secara signifikan dibandingkan dengan kelas lainnya (positif). Ketidakseimbangan semacam ini dapat menyebabkan masalah dalam pelatihan model machine learning, karena model cenderung lebih akurat dalam memprediksi kelas mayoritas (negatif) dan kurang efektif dalam mendeteksi kelas minoritas (positif).



Gambar 2 distribusi kelas diagnosis penyakit ginjal sesudah dan sebelum Smote

Oleh karena itu, teknik seperti SMOTE atau metode lain untuk menangani ketidakseimbangan data perlu diterapkan. Tujuannya adalah untuk menyeimbangkan distribusi kelas sehingga model dapat belajar dengan lebih baik dan menghasilkan prediksi yang lebih akurat untuk kedua kelas.

Dalam penelitian ini, ketidakseimbangan kelas dalam dataset pelatihan ditangani menggunakan teknik SMOTE. Sebelum penerapan SMOTE, distribusi kelas menunjukkan dominasi kelas mayoritas (negatif) sebesar 90,9%, sementara kelas minoritas (positif) hanya mencapai 9,1%. Ketidakseimbangan ini dapat menyebabkan model machine learning menjadi bias terhadap kelas mayoritas, sehingga mengurangi kemampuan deteksi terhadap kelas minoritas, yang merupakan fokus utama dalam penelitian ini. Setelah menerapkan SMOTE, dataset pelatihan mengalami penyeimbangan distribusi kelas. Hal ini dilakukan dengan menghasilkan sampel sintetis untuk kelas minoritas melalui interpolasi fitur dari k-nearest neighboring sampel minoritas. Akibatnya, jumlah sampel kelas minoritas meningkat dari 9,1% menjadi 50%, sehingga kedua kelas memiliki proporsi yang seimbang. Sebagai contoh, jika dataset asli terdiri dari 1.000 sampel (91 positif dan 909 negatif), setelah penerapan SMOTE, dataset pelatihan akan memiliki total 1.818 sampel.

Parameter yang di gunakan untuk menyamakan jumlah dataset mengunakan tekniks random_state bernilai 42. berfungsi sebagai seed generator untuk mengontrol stokastisitas dalam algoritma, memastikan proses acak dapat direproduksi secara identik di lingkungan berbeda. Dalam data mining, konsistensi ini krusial untuk validasi eksperimen, meminimalkan variabilitas hasil akibat inisialisasi acak berbasis waktu atau sumber sistem. Pemilihan nilai numerik (misal: 42) bersifat arbitrer namun terstandardisasi dalam praktik machine learning sebagai konvensi untuk menjamin konsistensi antar-eksekusi, meskipun tidak memiliki signifikansi matematis intrinsik. Angka 42

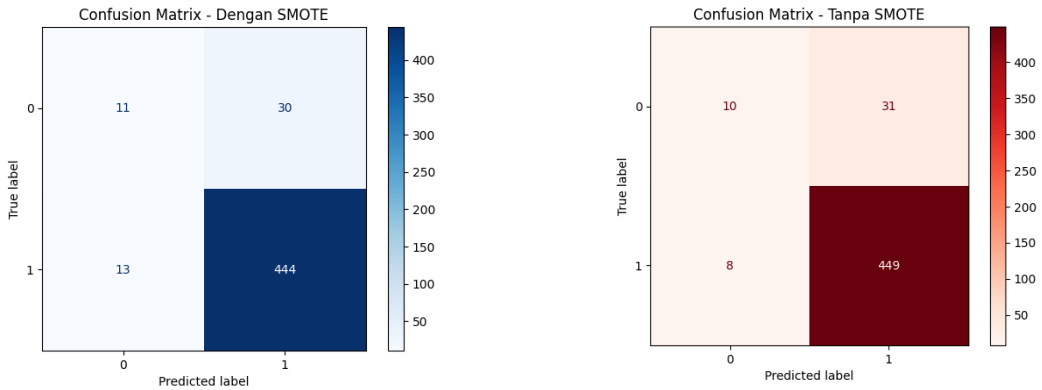
dipopulerkan oleh referensi budaya The Hitchhiker's Guide to the Galaxy, menjadi lelucon komunitas teknis yang sering diadopsi sebagai nilai default.

Pada implementasi SMOTE, random_state mengatur proses sintesis sampel minoritas melalui dua tahap:

1. Seleksi Instans Minoritas: Memilih titik data acak dari kelas minoritas.
2. Interpolasi Tetangga Terdekat: Menentukan k-nearest neighbors dari instans terpilih dan membangun sampel sintetis di ruang fitur.

Dengan penetapan random_state=42, kedua tahap ini menghasilkan urutan operasi identik tiap eksekusi, termasuk pemilihan tetangga dan koordinat interpolasi. Hal ini mencegah overfitting akibat variasi tak terkendali dalam augmentasi data sekaligus memfasilitasi perbandingan objektif antar-model. Dalam konteks penelitian, parameter ini menjadi komponen esensial untuk memastikan reproducibility dan memvalidasi dampak teknik resampling terhadap kinerja klasifikasi.

C. Evaluasi Model Gradient Boosting

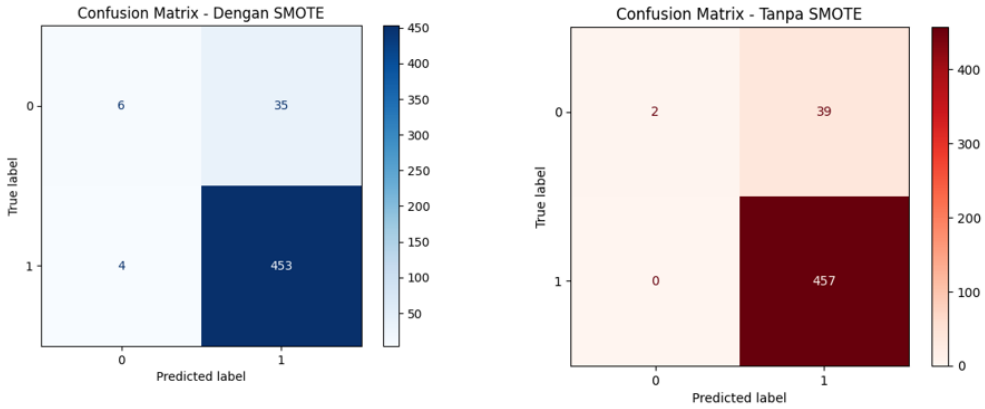


Gambar 3 perbandingan confusion matrix model gradient boosting

Dari gambar 3 yang merupakan confusion matrix hasil dari sebelum dan sesudah mengunakan metode smote dalam di liat bahwa Penerapan *Gradient Boosting* pada data asli menghasilkan akurasi validasi silang 93.28% dan AUC-ROC 0.8169, mengindikasikan bias implisit terhadap kelas mayoritas akibat *class distribution skew*. Ketidakmampuan model dalam menggeneralisasi pola kelas minoritas tercermin dari disparitas signifikan antara akurasi dan AUC-ROC, yang menandakan *generalization gap* pada prediksi kategori langka.

di Setelah augmentasi data menggunakan SMOTE, akurasi meningkat menjadi 95.13% dengan AUC-ROC melonjak ke 0.9949, mendekati kinerja sempurna. Sintesis sampel minoritas melalui interpolasi *k-nearest neighbors* memitigasi *overfitting* pada kelas dominan, memungkinkan *decision tree ensembles* dalam Gradient Boosting mengekstraksi *boundary decision* yang lebih presisi. Peningkatan AUC-ROC secara eksponensial mengonfirmasi efektivitas *resampling* dalam mengurangi *bias latent* dan meningkatkan *separability* ruang fitur. Hasil ini menegaskan bahwa ketidakseimbangan data secara struktural membatasi kapasitas model *tree-based* dalam *minority class recognition*, sementara SMOTE berperan sebagai *regularization implisit* dengan menyetarakan distribusi kelas, sehingga mengoptimalkan *loss function* selama fase *boosting*. Perbandingan ini dapat di liat pada gambar 4 sebagai alat perbandingan dalam pelaksanan metode smote.

D. Evaluasi Model Random Forest

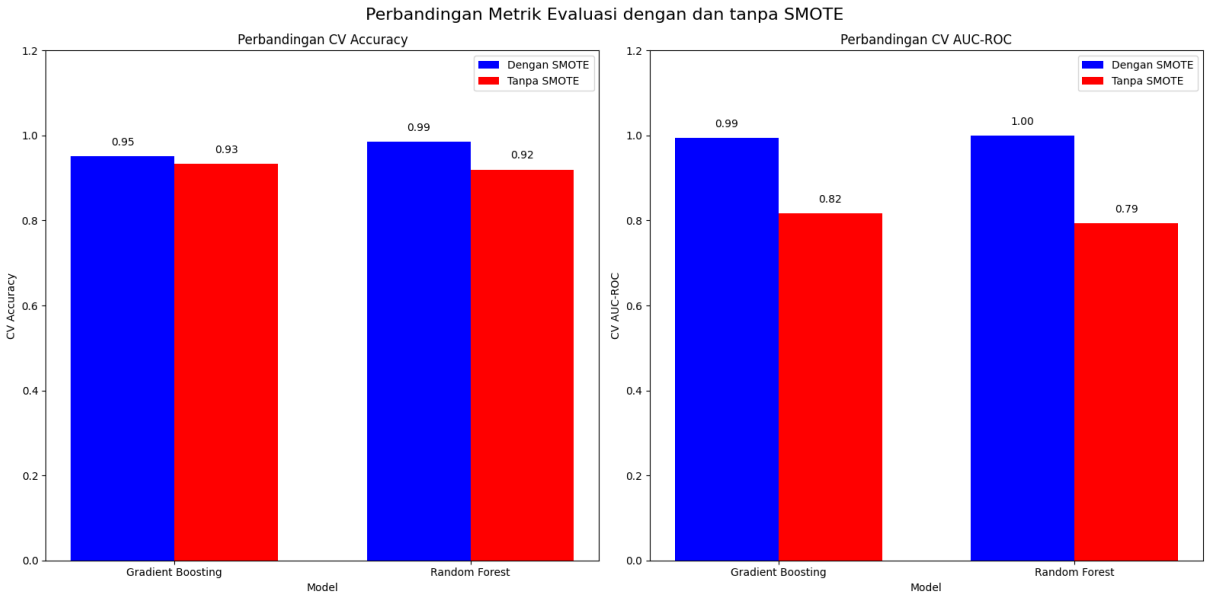


Gambar 4 perbandingan confusion matrix model Random Forest

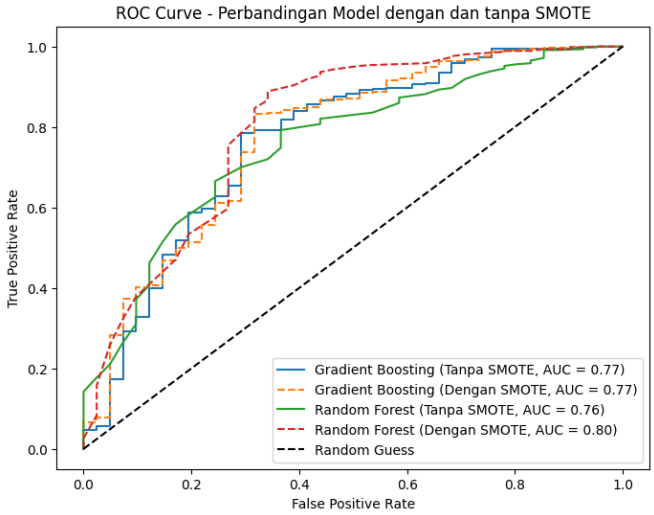
Penerapan *Random Forest* pada data asli menghasilkan akurasi validasi silang 91.99% dengan AUC-ROC 0.7928, mengindikasikan bias struktural terhadap kelas mayoritas akibat *imbalanced class distribution*. Rendahnya AUC-ROC mencerminkan ketidakmampuan *ensemble trees* dalam mengidentifikasi *decision boundary* kelas minoritas, yang tereduksi oleh dominansi sampel mayoritas dalam pembentukan *bootstrap aggregates*. Implementasi SMOTE meningkatkan akurasi ke 98.50% dan AUC-ROC ke 0.9992, menandakan transformasi signifikan dalam kemampuan generalisasi model. Generasi sampel sintetis melalui *k-nearest neighbors interpolation* mengatasi *variance imbalance* dengan memperkaya representasi kelas minoritas, memungkinkan *feature space* dipartisi lebih presisi selama konstruksi *decision trees*. Lonjakan AUC-ROC mendekati 1 mengkonfirmasi eliminasi *overlap* distribusi kelas, yang sebelumnya mengganggu optimasi *Gini impurity* dalam pemilihan *split points*. Hasil ini menggarisbawahi bahwa ketidakseimbangan data menghambat kapasitas *bagging* dalam menyeimbangkan *bias-variance trade-off*, sementara SMOTE berfungsi sebagai *data-centric regularization* yang memperkuat *minority class representation*. Peningkatan eksponensial AUC-ROC menunjukkan bahwa sintesis sampel tidak hanya menstabilkan *out-of-bag error*, tetapi juga memperdalam pemahaman model terhadap *manifold structure* data minoritas

E. Pembahasan

Gambar 5 membandingkan akurasi model Gradient Boosting dan Random Forest sebelum dan sesudah penerapan metode SMOTE (Synthetic Minority Over-sampling Technique). Hasil menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan akurasi kedua model. Pada Gradient Boosting, akurasi meningkat dari 93,28% menjadi 95,13%, sedangkan pada Random Forest, akurasi naik dari 91,99% menjadi 98,30%. Peningkatan ini menunjukkan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas, memungkinkan model untuk lebih baik dalam mengenali pola dari kedua kelas (positif dan negatif).



Gambar 5 Perbandingan Model Akurasi Dengan dan Tanpa Menggunakan Smote



Gambar 6 perbadingan nilai Area Under Curve sebelum dan sesudah smote

Pada gambar 6 menunjukkan grafik, Random Forest dengan SMOTE mencapai AUC (Area Under Curve) tertinggi sebesar 0.80, menunjukkan kemampuan terbaik dalam membedakan antara kelas positif dan negatif. Gradient Boosting memiliki AUC yang sama (0.77) baik dengan maupun tanpa SMOTE, mengindikasikan bahwa SMOTE tidak memberikan peningkatan signifikan pada model ini. Random Forest tanpa SMOTE memiliki AUC 0.76, sedikit lebih rendah dibandingkan dengan penerapan SMOTE. Garis *Random Guess* (tebakan acak) dengan AUC 0.5 berfungsi sebagai baseline, menunjukkan performa model yang tidak lebih baik dari tebakan acak. Secara keseluruhan, grafik ini mengilustrasikan bahwa Random Forest lebih responsif terhadap penerapan SMOTE dibandingkan Gradient Boosting, dengan peningkatan AUC yang lebih nyata. Hal ini menunjukkan bahwa SMOTE dapat efektif dalam meningkatkan kemampuan model tertentu, terutama dalam menangani ketidakseimbangan kelas.

IV. KESIMPULAN

Hasil penelitian menunjukkan bahwa SMOTE secara signifikan meningkatkan akurasi kedua model, dengan Gradient Boosting mencapai akurasi 95.28% dan Random Forest mencapai 99.30%. Hal ini membuktikan bahwa SMOTE efektif dalam mengatasi ketidakseimbangan kelas, memungkinkan model untuk mengenali pola dari kelas positif dan negatif secara lebih baik. Sebaliknya, Random Forest tetap mengungguli Gradient Boosting dalam kedua kondisi, dengan akurasi tertinggi mencapai 99.30%. Studi ini menekankan pentingnya penyeimbangan kelas dalam meningkatkan performa model, terutama untuk dataset yang tidak seimbang. Jika tujuan penelitian adalah meningkatkan keseimbangan antara Precision dan Recall, maka Gradient Boosting adalah pilihan yang lebih baik. Namun, jika fokus penelitian adalah meningkatkan akurasi dalam memprediksi instance positif, maka Random Forest lebih unggul.

REFERENCE

[1] P. Düsing *et al.*, "Vascular pathologies in chronic kidney disease: pathophysiological mechanisms and novel therapeutic approaches," *J. Mol. Med.*, vol. 99, no. 3, hal. 335–348, 2021, doi: 10.1007/s00109-021-02037-7.

[2] J. Portolés, L. Martín, J. J. Broseta, dan A. Cases, "Anemia in Chronic Kidney Disease: From Pathophysiology and Current Treatments, to Future Agents," *Front. Med.*, vol. 8, no. March, hal. 1–14, 2021, doi: 10.3389/fmed.2021.642296.

[3] V. K. Gliselda, "Diagnosis dan Manajemen Penyakit Ginjal Kronis (PGK)," *J. Med. Hutama*, vol. 2, no. 04 Juli, hal. 1135–1141, 2021.

[4] I. K. A. Loho, G. I. Rambert, dan M. F. Wowor, "Gambaran kadar ureum pada pasien penyakit ginjal kronik stadium 5 non dialisis," *J. e-Biomedik*, vol. 4, no. 2, hal. 2–7, 2016, doi: 10.35790/ebm.4.2.2016.12658.

[5] P. Putri dan A. T. Afandi, "Eksplorasi Kepatuhan Menjalani Hemodialisa Pasien Gagal Ginjal Kronik," *J. Keperawatan*, vol. 11, no. 2, hal. 37–44, 2022, doi: 10.47560/kep.v11i2.367.

[6] R. M. Hanna, A. Ferrey, C. M. Rhee, dan K. Kalantar-Zadeh, "Renal-Cerebral Pathophysiology: The Interplay Between Chronic Kidney Disease and Cerebrovascular Disease," *J. Stroke Cerebrovasc. Dis.*, vol. 30, no. 9, 2021, doi: 10.1016/j.jstrokecerebrovasdis.2020.105461.

[7] H. Yanai, H. Adachi, M. Hakoshima, dan H. Katsuyama, "Molecular biological and clinical understanding of the pathophysiology and treatments of hyperuricemia and its association with metabolic syndrome, cardiovascular diseases and chronic kidney disease," *Int. J. Mol. Sci.*, vol. 22, no. 17, 2021, doi: 10.3390/ijms22179221.

[8] G. Mulugeta, T. Zewotir, A. S. Tegegne, L. H. Juhar, dan M. B. Muleta, "Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia," *BMC Med. Inform. Decis. Mak.*, vol. 23, no. 1, hal. 1–17, 2023, doi: 10.1186/s12911-023-02185-5.

[9] A. K. Aggarwal, "Learning Texture Features from GLCM for Classification of Brain Tumor MRI Images using Random Forest Classifier," *Wseas Trans. Signal Process.*, vol. 18, no. April, hal. 60–63, 2022, doi: 10.37394/232014.2022.18.8.

[10] H. Tao, M. Habib, I. Aljarah, H. Faris, H. A. Afan, dan Z. M. Yaseen, "An intelligent evolutionary extreme gradient boosting algorithm development for modeling scour depths under submerged weir," *Inf. Sci. (Ny).*, vol. 570, hal. 172–184, 2021, doi: 10.1016/j.ins.2021.04.063.

[11] D. Elreedy, A. F. Atiya, dan F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, hal. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[12] S. Wang, Y. Dai, J. Shen, dan J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, hal. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[13] V. Werner de Vargas, J. A. Schneider Aranda, R. dos Santos Costa, P. R. da Silva Pereira, dan J. L. Victória Barbosa, "Imbalanced data preprocessing techniques for machine learning: a systematic mapping study," *Knowl. Inf. Syst.*, vol. 65, no. 1, hal. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.

[14] S. Gupta dan M. K. Gupta, "A comprehensive data-level investigation of cancer diagnosis on

imbalanced data," *Comput. Intell.*, vol. 38, no. 1, hal. 156–186, 2022, doi: 10.1111/coin.12452.
[15] Y. Fu, Y. Du, Z. Cao, Q. Li, dan W. Xiang, "A Deep Learning Model for Network Intrusion Detection with Imbalanced Data," *Electron.*, vol. 11, no. 6, hal. 1–13, 2022, doi: 10.3390/electronics11060898.
[16] Tedyyana, Agus, Osman Ghazali, and Onno W. Purbo. "Machine learning for network defense: automated DDoS detection with telegram notification integration." *Indonesian Journal of Electrical Engineering and Computer Science* 34.2 (2024): 1102..
[17] A. R. Salehi dan M. Khedmati, "A cluster-based SMOTE both-sampling (CSBBoost) ensemble algorithm for classifying imbalanced data," *Sci. Rep.*, vol. 14, no. 1, hal. 1–17, 2024, doi: 10.1038/s41598-024-55598-1.
[18] V. Nirmala, H. S. Shashank, M. M. Manoj, G. Satish Royal, dan J. Premaladha, "Skin Cancer Classification Using Image Processing with Machine Learning Techniques," *Intell. Data Anal. IoT, Blockchain*, hal. 1–15, 2023, doi: 10.1201/9781003371380-1.
[19] X. Wang *et al.*, "Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, hal. 1–14, 2021, doi: 10.1186/s12911-021-01471-4.
[20] M. Hanafy dan R. Ming, "Improving Imbalanced Data Classification in Auto Insurance by the Data Level Approaches," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, hal. 493–499, 2021, doi: 10.14569/IJACSA.2021.0120656.
[21] A. Gutiérrez-Gallego *et al.*, "Combination of Machine Learning Techniques to Predict Overweight/Obesity in Adults," *J. Pers. Med.*, vol. 14, no. 8, 2024, doi: 10.3390/jpm14080816.
[22] A. I. Putri *et al.*, "Implementation of K-Nearest Neighbors, Naïve Bayes Classifier, Support Vector Machine and Decision Tree Algorithms for Obesity Risk Prediction," *Public Res. J. Eng. Data Technol. Comput. Sci.*, vol. 2, no. 1, hal. 26–33, 2024, doi: 10.57152/predatecs.v2i1.1110.
[23] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere, dan S. Hussain, "Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks," *Appl. Sci.*, vol. 13, no. 6, 2023, doi: 10.3390/app13064006.
[24] Asniar, N. U. Maulidevi, dan K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, hal. 3413–3423, 2022, doi: 10.1016/j.jksuci.2021.01.014.
[25] A. Anggrawan, H. Hairani, dan C. Satria, "Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE," *Int. J. Inf. Educ. Technol.*, vol. 13, no. 2, hal. 289–295, 2023, doi: 10.18178/ijiet.2023.13.2.1806.
[26] H. Hairani, A. Anggrawan, dan D. Priyanto, "Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link," *Int. J. Informatics Vis.*, vol. 7, no. 1, hal. 258–264, 2023, doi: 10.30630/joiv.7.1.1069.
[27] N. Koklu dan S. A. Sulak, "Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities," *Sinop Üniversitesi Fen Bilim. Derg.*, vol. 9, no. 1, hal. 217–239, 2024, doi: 10.33484/sinopfbd.1445215.
[28] A. Frattini, I. Bianchini, A. Garzonio, dan L. Mercuri, "Financial Technical Indicator and Algorithmic Trading Strategy Based on Machine Learning and Alternative Data," *Risks*, vol. 10, no. 12, hal. 1–24, 2022, doi: 10.3390/risks10120225.
[29] M. Amjad, I. Ahmad, M. Ahmad, P. Wróblewski, P. Kamiński, dan U. Amjad, "Prediction of Pile Bearing Capacity Using XGBoost Algorithm: Modeling and Performance Evaluation," *Appl. Sci.*, vol. 12, no. 4, 2022, doi: 10.3390/app12042126.
[30] M. Iqbal, W. S. Dharmawan, dan R. Septian, "Journal of Computer Networks , Architecture and High Performance Computing Prediction of Obesity Categories Based on Physical Activity Using Machine Learning Algorithms Journal of Computer Networks , Architecture and High Performance Computing," *J. Comput. Networks, Archit. High Perform. Comput.*, vol. 6, no. 3, hal. 1025–1034, 2024.
[31] S. Bakheet dan A. Al-Hamadi, "Automatic detection of COVID-19 using pruned GLCM-Based texture features and LDCRF classification," *Comput. Biol. Med.*, vol. 137, no. June, hal. 104781, 2021, doi: 10.1016/j.compbiomed.2021.104781.
[32] Y. Wang, S. Ye, Z. Xu, Y. Chu, J. Zhang, dan W. Yu, "Research on Sleep Staging Based on Support Vector Machine and Extreme Gradient Boosting Algorithm," *Nat. Sci. Sleep*, vol. 16, hal. 1827–1847, 2024, doi: 10.2147/NSS.S467111.