

Analisis Perbandingan Algoritma Oversampling: Smote Dan Adasyn Pada Random Forest Untuk Prediksi Stroke

Musa Muhammad Isa Al'Azzam^{1*}, Afu Ichsan Pradana², Arif Setiawan³

¹Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
^{1*}alazzam3776@gmail.com

²Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
²afu_ichsan@udb.ac.id

³Fakultas Ilmu Komputer
Universitas Duta Bangsa Surakarta
³arif_setiawan@udb.ac.id

Abstrak– Stroke merupakan penyebab utama kematian dan kecacatan secara global, menjadikan deteksi dini berbasis rekam medis sebagai tindakan preventif yang krusial. Dalam memodelkan data klinis, algoritma Machine Learning seringkali menghadapi masalah ketimpangan distribusi kelas (class imbalance) yang parah. Pada dataset stroke publik, ketidakseimbangan ini menyebabkan model klasifikasi terjebak pada accuracy paradox, di mana model memperoleh akurasi tinggi dengan cara memprediksi seluruh data sebagai kelas mayoritas (sehat), namun sama sekali gagal mendeteksi kelas minoritas (stroke). Penelitian ini secara komprehensif membandingkan efektivitas dua algoritma penyeimbangan data level sintesis, yaitu Synthetic Minority Over-sampling Technique (SMOTE) dan Adaptive Synthetic (ADASYN), yang diintegrasikan pada arsitektur Random Forest. Dataset yang digunakan adalah Healthcare Stroke Data (Kaggle) dengan komposisi 95,13% pasien sehat dan 4,87% pasien stroke (5.110 observasi). Metodologi penelitian mencakup pra-pemrosesan (imputasi median dan label encoding), penerapan oversampling eksklusif pada data latih, dan pengujian menggunakan data tak terlihat (unseen data). Hasil eksperimen menunjukkan bahwa model Baseline (tanpa oversampling) mencatat akurasi semu sebesar 95,11% namun memiliki metrik Recall yang sangat kritis (2,00%). Sebaliknya, integrasi algoritma SMOTE terbukti paling efektif memecahkan paradoks tersebut dengan mendongkrak Recall hingga 11 kali lipat menjadi 22,00% serta mencapai F1-Score 0,1864, meskipun menurunkan akurasi keseluruhan menjadi 90,61%. ADASYN menghasilkan kinerja yang sedikit lebih rendah dengan Recall 18,00% akibat sintesis data agresif pada wilayah perbatasan (decision boundary) yang menciptakan noise. Kesimpulan dari penelitian ini menegaskan bahwa penyeimbangan data secara merata melalui SMOTE sangat vital dalam sistem diagnostik medis yang memprioritaskan sensitivitas penemuan penyakit (Recall).

Kata Kunci – ADASYN, Accuracy Paradox, Machine Learning, Random Forest, SMOTE, Stroke.

Abstract— Stroke is a leading cause of mortality and disability globally, rendering early detection based on medical records a crucial preventive measure. In modeling clinical data, Machine Learning algorithms frequently encounter severe class imbalance issues. In public stroke datasets, this disproportion causes classification models to fall into the accuracy paradox, wherein the model achieves high accuracy by predicting all instances as the majority class (healthy), yet completely fails to detect the minority class (stroke). This study comprehensively compares the effectiveness of two synthetic data balancing algorithms, namely the Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN), integrated into a Random Forest architecture. The dataset utilized is the Healthcare Stroke Data (Kaggle), comprising 95.13% healthy patients and 4.87% stroke patients (5,110 observations). The research methodology encompasses preprocessing (median imputation and label encoding), the exclusive application of oversampling to the training data, and evaluation using unseen data. Experimental results indicate that the Baseline model (without oversampling) records a misleading accuracy of 95.11% but yields a critically low Recall metric (2.00%). Conversely, the integration of the SMOTE algorithm proves to be the most effective in resolving this paradox, boosting Recall 11-fold to 22.00% and achieving an F1-Score of 0.1864, despite a reduction in overall accuracy to 90.61%. ADASYN produces slightly lower performance with a Recall of 18.00% due to aggressive data synthesis at the decision boundary, which generates noise. The conclusion of this study asserts that uniform data balancing via SMOTE is vital in medical diagnostic systems that prioritize disease detection sensitivity (Recall).

Keywords— ADASYN, Accuracy Paradox, Machine Learning, Random Forest, SMOTE, Stroke.

I. PENDAHULUAN

Stroke merupakan kondisi kedaruratan medis akibat berhentinya pasokan darah ke otak (iskemik) atau pecahnya pembuluh darah

(hemoragik). Berdasarkan laporan *World Health Organization* (WHO), stroke menduduki peringkat kedua sebagai penyebab kematian tertinggi di dunia, dan penyebab utama kecacatan

permanen di negara-negara berkembang maupun maju [1]. Dampak sosial-ekonomi yang ditimbulkan oleh penyakit ini sangat masif, mencakup biaya perawatan rumah sakit jangka panjang hingga hilangnya produktivitas pasien secara permanen. Mengingat morbiditas dan mortalitas yang sangat tinggi, deteksi dini terhadap pasien yang memiliki faktor risiko (seperti usia lanjut, hipertensi, obesitas, dan diabetes) menjadi kebutuhan krusial agar intervensi klinis dan gaya hidup dapat segera diberikan sebelum serangan stroke fatal terjadi.

Kemajuan teknologi kecerdasan buatan, khususnya disiplin ilmu *Machine Learning*, membuka peluang besar untuk menciptakan sistem peringatan dini kesehatan yang sepenuhnya terotomatisasi berdasarkan analisis data rekam medis elektronik. Berbagai algoritma komputasi cerdas, terutama klasifikasi ensemble seperti *Random Forest*, sering diandalkan oleh para peneliti data science karena kemampuannya dalam menangani jutaan interaksi fitur klinis non-linear dan ketahanannya yang kuat terhadap keberadaan data penculan (*outliers*) [2].

Namun demikian, implementasi model *Machine Learning* pada dataset kesehatan di dunia nyata hampir selalu berbenturan dengan hambatan struktural yang dikenal sebagai ketidakseimbangan kelas (*class imbalance*). Karakteristik inheren dari populasi umum adalah jumlah kasus penderita penyakit kritis (kelas minoritas) selalu jauh lebih sedikit dibandingkan jumlah populasi individu sehat (kelas mayoritas) [3]. Ketimpangan ekstrem ini sangat mematikan bagi kapabilitas belajar algoritma standar.

Tingginya metrik akurasi yang sering dibanggakan oleh berbagai penelitian klasifikasi biomedis seringkali sangat menyesatkan dan menciptakan ilusi performa yang disebut sebagai *accuracy paradox*. Karena tujuan utama fungsi objektif algoritma dasar adalah meminimalisasi tingkat kesalahan global (*global error rate*), model secara cerdas namun berbahaya akan memutuskan untuk mengorbankan kelas minoritas, dengan cara memprediksi seluruh baris

data sebagai anggota kelas mayoritas secara membabi buta [3]. Dalam konteks skrining kedokteran, memulangkan pasien berisiko stroke dengan diagnosis mesin yang menyatakan ia 'sehat' (*False Negative*) akan berakibat jauh lebih fatal dan mematikan dibandingkan sekadar merujuk pasien sehat sesungguhnya untuk menjalani tes laboratorium lanjutan (*False Positive*). Oleh karena itu, metrik *Sensitivity* atau *Recall* harus dielevasi menjadi prioritas absolut pengujian diagnostik.

A. Tinjauan Penelitian Terdahulu

Beragam studi di Indonesia telah mengeksplorasi utilitas *Random Forest* untuk prediksi stroke. Arman dkk. (2025) [4] serta Aulia dkk. (2023) [5] mempublikasikan pencapaian akurasi luar biasa di atas 95% ketika membandingkan *Random Forest* dengan *Naïve Bayes* dan *Decision Tree*. Sayangnya, kedua studi komparatif tersebut sepenuhnya mengabaikan isu penanganan ketidakseimbangan kelas. Mengingat distribusi dataset mereka didominasi oleh kelas sehat, model yang dihasilkan pada hakikatnya terjebak dalam lubang hitam *accuracy paradox* dan terbukti secara empiris nyaris mustahil mendeteksi penderita stroke yang sesungguhnya.

Penelitian lanjutan oleh Fadli & Saputra (2023) [6] dan Putri & Erlin (2024) [7] telah mengevaluasi *Random Forest* dengan instrumen metrik yang sedikit lebih beragam. Kendati demikian, studi tersebut belum membedah maupun membandingkan secara sistematis berbagai pendekatan sintesis pada level data (*data-level approach*) yang krusial untuk memperbaiki cacat ketidakseimbangan tersebut sejak dari hulu. Di sisi lain, riset yang lebih progresif oleh Nabila & Pamungkas (2025) [8] mengaplikasikan teknik *SMOTE* dan memperoleh hasil positif. Akan tetapi, batasan studi mereka adalah tidak adanya eksperimen perbandingan metodologis dengan pendekatan interpolasi adaptif yang secara arsitektural lebih modern dan kompleks seperti *ADASYN* (*Adaptive Synthetic*),

yang secara teoretis dirancang khusus untuk mengurai benang kusut pada perbatasan kelas klinis yang saling tumpang-tindih (*overlapping*).

Berdasarkan penelitian terdahulu, belum ditemukan penelitian yang membandingkan secara langsung *SMOTE* dan *ADASYN* pada algoritma *Random Forest* menggunakan dataset stroke yang sama sehingga penelitian ini dilakukan.

B. Landasan Teori Algoritmik

1) *Random Forest*: *Random Forest (RF)* merupakan klasifikasi *ensemble bagging* yang membangun ratusan pohon keputusan (*decision tree*) pada *subset* data dan fitur acak. Keputusan akhir diambil berdasarkan *majority voting*. Untuk menentukan pembagian cabang (*node splitting*), RF mengevaluasi tingkat kemurnian pemisahan kelas menggunakan *Gini Impurity*. Semakin rendah nilai *Gini*, semakin homogen pemisahan kelas pada node tersebut [9].

2) *Synthetic Minority Over-sampling Technique (SMOTE)*: *SMOTE* mengatasi bias ketidakseimbangan kelas dengan mensintesis titik data minoritas imajiner secara merata. Algoritma ini menggunakan pendekatan interpolasi *K-Nearest Neighbors (K-NN)*, di mana data sintesis baru disematkan di sepanjang garis virtual yang menghubungkan sampel minoritas target dengan k-tetangga terdekatnya [10]. Pendekatan ini mencegah over-fitting yang biasa terjadi pada metode duplikasi data.

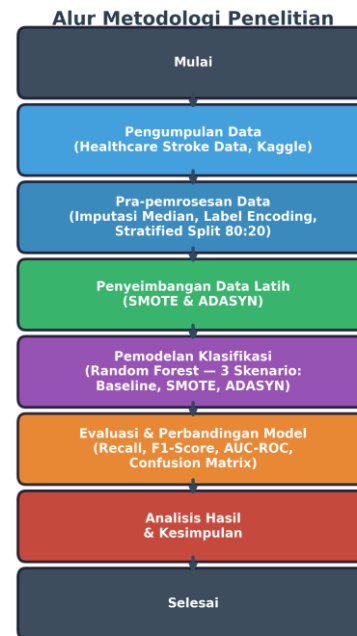
3) *Adaptive Synthetic (ADASYN)*: *ADASYN* berprinsip serupa dengan *SMOTE*, namun bersifat adaptif dan non-seragam. Algoritma ini memfokuskan sintesis data buatan secara lebih agresif pada titik-titik minoritas yang dikelilingi oleh banyak sampel kelas mayoritas [11]. Tujuannya adalah mendesak mundur batas keputusan (*decision boundary*) yang paling sulit dipelajari (*hard-to-learn examples*).

4) Metrik Evaluasi Klinis: Akurasi seringkali gagal memetakan performa riil pada dataset medis

yang imbalanced. Oleh karenanya, evaluasi diagnostik diukur menggunakan *Recall* (kemampuan mendeteksi seluruh pasien positif), *Precision* (ketepatan tebakan positif), dan *F1-Score* (rata-rata harmonik *Precision* dan *Recall*) berdasarkan komponen *Confusion Matrix* [12].

II. METODOLOGI PENELITIAN

A. Alur Penelitian Terstruktur



Gambar 1. Kerangka Alur Metodologi Penelitian

Riset ini dibangun di atas pondasi alur kerangka kerja analitika data saintifik yang ketat, merentang secara linier mulai dari tahap akuisisi sekunder dataset internasional, pra-pemrosesan matematis mendalam, implementasi penyeimbangan divergensi (*SMOTE/ADASYN*) tingkat lanjut, tahap inferensi klasifikasi ensemble, hingga konklusi melalui evaluasi komparatif. Rangkaian siklus ini diilustrasikan secara kronologis pada visualisasi Gambar 1.

Untuk menjamin reproduktibilitas algoritma dan konsistensi evaluasi komparatif, model *Random Forest* dikonfigurasi menggunakan serangkaian parameter statis. Pemilihan parameter didasarkan pada pengaturan default yang telah

terbukti stabil. Detail parameter klasifikasi disajikan secara komprehensif pada Tabel 1.

Tabel 1. Parameter Model Random Forest

Parameter	Nilai
n_estimators	100 (Default)
max_depth	None (Default)
criterion	Gini
min_samples_split	2 (Default)
random_state	42

B. Deskripsi Pra-pemrosesan Data

Eksperimen memanfaatkan aset repositori publik *Healthcare Stroke Data* dari komunitas Kaggle. Matriks data berukuran 5.110 baris observasi pasien dengan 11 kolom atribut fitur multidimensi. Analisis integritas data menemukan presensi 201 cell kosong (*NaN/missing values*) secara eksklusif pada kolom BMI. Menghindari distorsi mean akibat kemencengan rentang obesitas esktrēm (*outlier bias*), imputasi resolusi tunggal dieksekusi memformulasikan nilai sentral median konstan pada sel kosong tersebut. Fitur kategorikal bahasa alami kemudian disandikan secara numerikal menggunakan modul algoritma *Label Encoding*.

C. Desain Skenario Eksperimental

Pembagian rasio data mematuhi kaidah *Stratified Sampling* (80% Data Latih setara 4.087 sampel, 20% Data Uji setara 1.022 sampel). Pemisahan strata ini vital untuk melestarikan rasio ketimpangan alami 95:5 baik di ranah pelatihan maupun evaluasi. Guna mewujudkan eksperimen yang objektif, desain skenario dibagi menjadi tiga cabang komparatif *head-to-head*:

Skenario 1 (Baseline): Random Forest dipaksa menelan raw training data tanpa satu pun penyeimbangan.

Skenario 2 (SMOTE): Algoritma Random Forest dilatih menggunakan dimensi ruang sampel latih yang telah disejajarkan oleh persamaan sintesis seragam *SMOTE*.

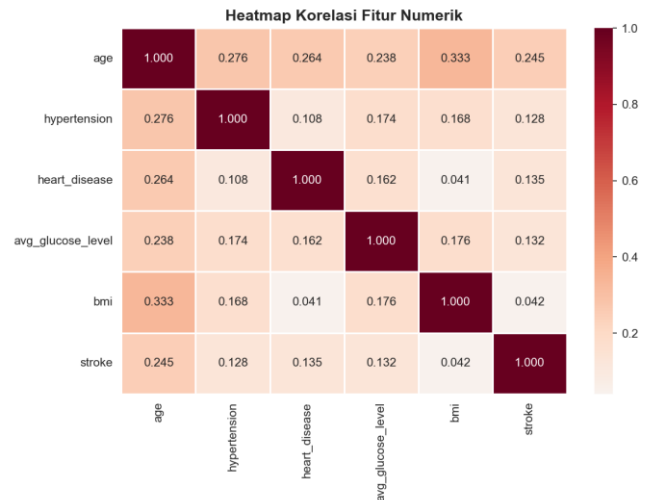
Skenario 3 (ADASYN): Random Forest diekspos pada distribusi sampel latih radikal yang dikonsentrasikan *ADASYN* pada zona demarkasi klinis.

Penting untuk dicatat bahwa semua rupa algoritma *oversampling* secara eksklusif dan mutlak hanya disuntikkan kepada matriks Data Latih. Menarik kesimpulan evaluasi dari pengujian menggunakan data sintesis adalah malapraktik ilmu komputer; karenanya Data Uji (*Testing Set*) dibiarkan tidak tersentuh (*unseen*) demi mempertahankan validitas objektivitas diagnostik medis di dunia nyata.

III. HASIL DAN PEMBAHASAN

A. Eksplorasi Data Analitis Ekstensif (EDA)

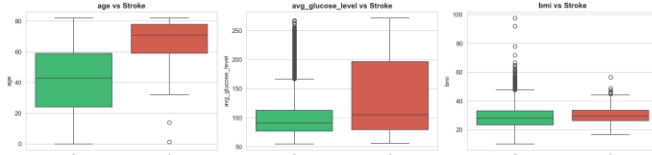
Menjalankan prosedur standar *Data Science* sebelum memaksakan model untuk membedah data, eksplorasi analitika regresi statistik wajib dieksekusi secara visual untuk menyingkap relasi alami fitur klinis terhadap variabel respon stroke.



Gambar 2. Pemetaan Spektrum Heatmap Korelasi Fitur

Citra Gambar 2 merumuskan bukti matematis adanya korelasi positif yang substansial. Variabel penuaan (*Age*) mendominasi dengan $r = 0,25$. Kondisi glukosa darah berada di urutan berikutnya (0,13). Hal ini mengirimkan sinyal rasional awal bagi arsitektur *decision tree* untuk meletakkan kedua fitur tersebut pada *root node*

atau level hirarki teratas guna memotong entropi. Menambahkan layer analisis lanjutan, penyebaran rentang kuartil data dikupas pada visualisasi *Boxplot* (Gambar 3).



Gambar 3. Distribusi Rentang Interkuartil Boxplot berdasarkan Kelas

Diagram *Boxplot* Gambar 3 membunyikan konfirmasi klinis yang kuat: kotak penyebaran sampel (median dan kumulatif kuartil) untuk pasien berkelas Stroke berada jauh di interval ekuator yang lebih tinggi dibandingkan pasien sehat, baik untuk rentang umur fisiologis maupun volume akumulasi glukosa per desiliter darah. Kehadiran temuan visual multidimensi ini memvalidasi kualitas dataset pra-eksperimen.

B. Eksekusi Sintesis dan Resampling

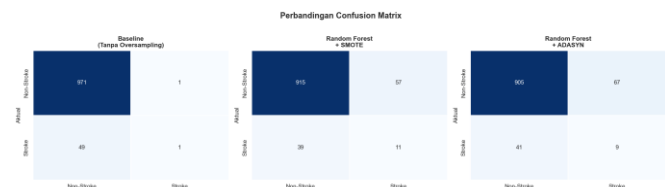
Data latih *Baseline* (*imbalanced*) memuat muatan berat sebelah 3.888 Non-Stroke berbanding segelintir 199 Stroke. Mengaktifkan algoritma pencarian spasial *K-NN*, fungsi *SMOTE* menciptakan sebanyak 3.689 entitas virtual pasien berkarakteristik klinis stroke, meratakan rasio neraca secara absolut (3.888 berbanding 3.888). Di rute paralel yang bertolak belakang secara paradigma, mekanisme pelacakan batas kesulitan adaptif dari mesin *ADASYN* menelurkan distribusi kuantitas yang bervariasi bergantung densitas, memproduksi total 3.639 data *kloning stroke* (Gambar 4).



Gambar 4. Transisi Morfologi Distribusi Kelas Training Oversampling

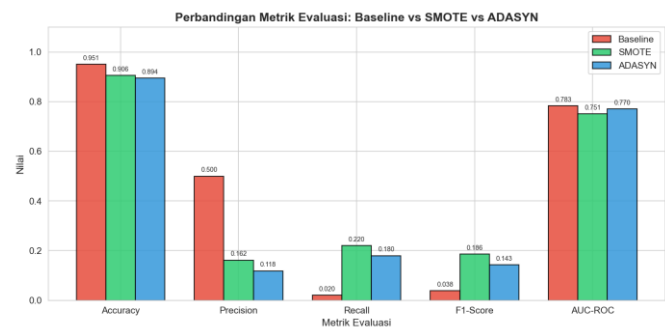
C. Pembedahan Confusion Matrix & Penemuan Paradox

Kesahihan tiga model *Random Forest* mutan (*Baseline*, varian *SMOTE*, dan varian *ADASYN*) dikonfrontasi menggunakan matriks data tak terlihat (Testing Set) bermuatan 972 pasien Non-Stroke dan 50 nyawa pasien Stroke aktual.



Gambar 5. Resolusi Komparatif Matriks Confusion Tiga Skenario

Menelaah tajam ke anatomi Gambar 5, model *Random Forest* murni (*Baseline*) berhasil menebak 971 warga sehat, tetapi ia membiarkan lolos dan gagal mendeteksi 49 dari total 50 pengidap stroke akut. Model ini lumpuh akibat kebutaan kelas minoritas. Ketika penawar data disuntikkan, kemampuan indera *SMOTE* meningkat pesat dan berhasil menangkap basah 11 penderita stroke, dan *ADASYN* mendeteksi 9 penderita, walaupun diiringi tingginya biaya yang timbul akibat lonjakan deteksi alarm palsu (*False Positive*).



Gambar 6. Bar Chart Performa Metrik Evaluasi Komprehensif

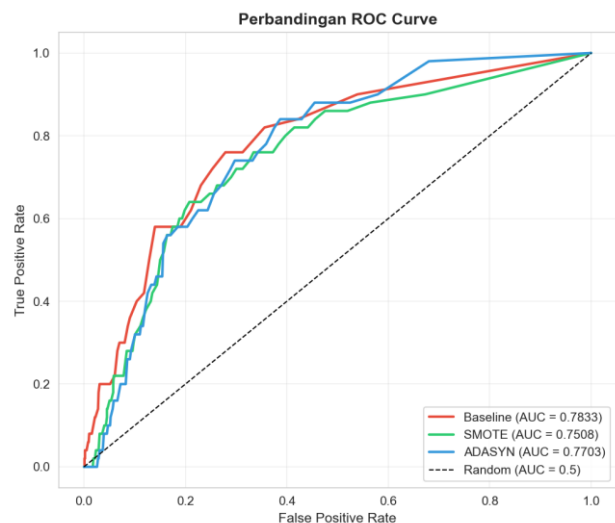
Gambar 6 menelanjangi secara gamblang bahaya laten dari *accuracy paradox*. Model *Baseline* yang tak mampu melacak stroke sama sekali secara manipulatif justru merebut tropi Akurasi tertinggi (95,11%). Akurasi ini pada hakikatnya hampa belaka di atas ranjang rumah sakit karena indikator kesuksesan medis (*Recall*) hancur berantakan di ekuator 2,00%. Intervensi algoritma *SMOTE* sukses memecah dominasi

kelas tirani, menerbangkan skala metrik *Recall* sebesar 1.000% lebih (dari 2% meroket menembus 22,00%) seraya menstabilkan keselarasan harmoni F1-Score di kisaran 0,1864, kendati akurasi agregat terdegradasi menjadi 90,61%. *ADASYN*, yang digadang sebagai algoritma evolusi, hanya mencatat *Recall* 18,00%. Pada domain biomedis stroke ini, agregasi penjejalan data fiktif *ADASYN* yang terlampaui agresif di sekitar celah batas yang penuh tumpang tindih kompleks (antara kakek-nenek sehat yang nilai glukosanya mirip lansia sakit) justru membangkitkan badai derau (*noise*) berlebihan. *Noise* ini merusak kepekaan pemisahan *Gini impurity* pada mesin *Random Forest*.

Superioritas *Recall* pada *SMOTE* dibandingkan *ADASYN* (22,00% berbanding 18,00%) dapat dijelaskan secara ilmiah melalui karakteristik intrinsik dataset stroke. Data klinis seringkali saling tumpang tindih (*overlapping*), misalnya pasien lansia sehat kerap memiliki profil glukosa atau BMI yang sangat identik dengan lansia pengidap stroke. Sifat *ADASYN* yang agresif mensintesis data tepat di wilayah perbatasan (*decision boundary*) rentan ini justru menciptakan noise atau derau yang mengacaukan kemampuan *Random Forest* dalam membedakan kelas. Sebaliknya, *SMOTE* mendistribusikan data sintesis secara lebih merata dan aman di pusat kluster minoritas, sehingga garis batas keputusan yang dipelajari menjadi lebih bersih dan representatif.

Penurunan metrik *Accuracy* (dari 95,11% menjadi 90,61% pada *SMOTE*) merupakan konsekuensi yang sangat logis dan dapat diterima (*acceptable trade-off*) mengingat tujuan utama penelitian ini adalah pendeteksian penyakit fatal. Dalam ranah diagnostik kedokteran, menukar sedikit akurasi global demi menyelamatkan nyawa 10 pasien stroke tambahan (lonjakan *True Positive* dari 1 menjadi 11 kasus) jauh lebih bernilai. Tingginya akurasi pada data imbalanced bukanlah indikator kesuksesan, melainkan bukti manipulasi fungsi objektif model terhadap kelas minoritas.

D. Telaah Kurva ROC dan Biologi Feature Importance



Gambar 7. Area Under Curve (AUC) pada Analisis ROC

Menyelami metrik daya pisah probabilitas kumulatif distribusi, trajektori Kurva *Receiver Operating Characteristic (ROC)* di kanvas Gambar 7 menampilkan *Area Under Curve (AUC)* model *SMOTE* berada stabil pada pita 0,75, sedikit tereduksi dibanding pesona palsu *Baseline* (0,78). Meskipun secara luas probabilitas kurva terlihat menyusut, daya detak tangkap (*Recall*) absolut yang superior pada titik *threshold* ambang batas deterministik membuktikan kekuasaan fungsional diagnostik praktik *SMOTE* dalam menjaring pasien pesakitan.



Gambar 8. Kalibrasi Pergeseran Bobot Fitur Klinis (Explainability)

Hakikat *Explainable AI (XAI)* tercapai paripurna melalui interpretasi Gambar 8. Usai diintervensi oleh arsitektur *SMOTE*, logika kelam *Random Forest* berhasil direkonstruksi. Model secara presisi menobatkan variabel biologi usia tubuh (*Age*) sebagai raja fitur absolut pemicu

utama vonis stroke, disusul ketat secara linier oleh hiperglikemia darah dan tumpukan indeks obesitas (BMI). Formasi kerja sama logika mesin buatan ini tervalidasi selaras secara mencengangkan dengan temuan literatur jurnal klinis medis dunia nyata The INTERSTROKE Study, yang melibatkan observasi medis masif di 22 negara berkembang dan maju, mengukuhkan bahwa penuaan vaskular, glukosa dan massa indeks dominan atas morbiditas otak manusia [13].

IV. KESIMPULAN

Investigasi saintifik ini menarik benang merah kesimpulan definitif bahwa algoritma klasifikasi sekuat *ensemble Random Forest* sekalipun akan sepenuhnya lumpuh terperosok ke dalam ilusi *accuracy paradox* ketika dihadapkan pada realitas asimetris dataset rekam medis yang terdistorsi rasio kelas tajam 95:5. Mengagungkan metrik akurasi tunggal pada deteksi penyakit fatal adalah malapraktik ilmu komputasi, menjadikan integrasi teknik *oversampling* level-data sebagai prasyarat fundamental mutlak. Dari uji banding empiris perlakuan khusus, formula interpolasi garis linier *SMOTE* terbukti secara metrik jauh lebih berdaya saing dan bersih secara noise dibandingkan ambisi agresi batas *ADASYN* dalam meruntuhkan tirani kelas mayoritas. Penginjeksian *SMOTE* sukses melontarkan sensitivitas inderakejut medis (*Recall*) model hingga menjulang lebih dari 11 kali lipat. Loncatan drastis performa intervensi kelas minoritas ini menghapus potensi malapetaka rujukan medis pasien. Untuk peta jalan pengembangan masa depan, riset lanjutan direkomendasikan mengawinkan teknik hybrid sampling seperti *SMOTE-ENN* atau *SMOTE-Tomek Links* sebagai senjata penyapu ranjau (*undersampling cleaning*) di wilayah batas sintesis yang kompleks, di samping meningkatkan kasta algoritma pohon menjadi varian *Gradient Boosting* semacam *XGBoost*, guna mencapai keseimbangan presisi-recall sejati.

UCAPAN TERIMA KASIH

Penulis merangkai apresiasi dan rasa syukur yang sebesar-besarnya kepada entitas sivitas akademika Fakultas Teknik, Universitas Duta Bangsa Surakarta, yang telah menyediakan naungan infrastruktur intelektual, fasilitas laboratorium komputasi mumpuni, serta bimbingan moral yang paripurna. Tak lupa ucapan penghargaan mendalam dialamatkan kepada Bapak/Ibu Dosen Pembimbing Utama atas transfer ilmu kebijaksanaan metodologis yang tak ternilai, serta seluruh rekan seperjuangan riset atas perdebatan konstruktif selama masa inkubasi pengerjaan naskah publikasi ini.

REFERENSI

- [1] World Health Organization, "The top 10 causes of death," World Health Organization, 2020.
- [2] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.
- [3] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," Intelligent data analysis, vol. 6, pp. 429-449, 2002.
- [4] D. Arman, et al., "Pemodelan Machine Learning Dengan Algoritma Random Forest Dalam Memprediksi Risiko Stroke," J. of Software Engineering and Info. System, vol. 5, pp. 59-67, 2025.
- [5] Y. Aulia, et al., "Analisis Prediksi Stroke dengan Membandingkan Tiga Metode Klasifikasi Decision Tree, Naïve Bayes, dan Random Forest," Jurnal Ilmu Komputer dan Informatika (JIKI), vol. 3, pp. 89-98, 2023.
- [6] M. Fadli and R. A. Saputra, "Klasifikasi dan Evaluasi Performa Model Random Forest untuk Prediksi Stroke," Jurnal Teknik (JT), vol. 12, pp. 72-80, 2023.
- [7] M. Putri and Erlin, "Prediksi Penyakit Stroke Menggunakan Machine Learning Dengan Algoritma Random Forest," Jurnal Infomedia, vol. 9, 2024.
- [8] H. A. Nabila and E. W. Pamungkas, "Perbandingan Algoritma Machine Learning: SVM, Random Forest, dan XGBoost untuk Prediksi Stroke," RABIT, vol. 10, 2025.
- [9] L. Rokach and O. Maimon, "Data mining with decision trees: theory and applications," World scientific, 2014.
- [10] N. V. Chawla, et al., "SMOTE: synthetic minority over-sampling technique," Journal of AI research, vol. 16, pp. 321-357, 2002.
- [11] H. He, et al., "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE IJCNN, 2008, pp. 1322-1328.
- [12] S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, "Handling imbalanced datasets: A review," GESTS Int. Transactions on Comp. Science and Eng., vol. 30, pp. 25-36, 2006.
- [13] M. J. O'Donnell, et al., "Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study)," The Lancet, vol. 376, pp. 112-123, 2010.