Evaluasi Model Klasifikasi Penyakit Stroke: Pendekatan Machine Learning Klasik, Deep Learning, dan Explainable AI

First Author1, Second Author2, ..., Fifth Author5

1,2,3,4,5Author’s affiliation (2 - 5 authors)

first\_or\_corresponding.author@first.edu

**Abstract - This study aims to analyze the impact of the Synthetic Minority Oversampling Technique (SMOTE) on improving the performance of machine learning and deep learning models in stroke risk classification, a task often challenged by class imbalance. A quantitative experimental design was employed using a public dataset that underwent preprocessing steps including duplicate removal, missing value handling, correlation-based feature selection, and normalization. The evaluated models consisted of Logistic Regression and Random Forest for machine learning, as well as Sequential and TabNet for deep learning. Model performance was assessed using accuracy, precision, recall, and F1-score metrics. Model interpretation was conducted through Explainable Artificial Intelligence (XAI) using SHAP for global analysis and LIME for local analysis. The results indicate that applying SMOTE significantly enhances model performance, particularly in detecting the minority “Stroke” class, with Random Forest after SMOTE achieving 97% accuracy and optimal precision–recall balance. These findings support the hypothesis that data balancing techniques can improve low-prevalence case detection, offering practical contributions to the development of transparent and reliable AI-based health screening systems.**

**Keywords: SMOTE, machine learning, deep learning, stroke, XAI**

**Abstrak - Penelitian ini bertujuan menganalisis pengaruh *Synthetic Minority Oversampling Technique* (SMOTE) terhadap peningkatan kinerja model *machine learning* dan *deep learning* dalam klasifikasi risiko stroke, yang kerap mengalami masalah ketidakseimbangan kelas. Desain penelitian menggunakan pendekatan kuantitatif eksperimental dengan pemanfaatan dataset publik yang telah melalui tahap pra-pemrosesan, termasuk pembersihan data duplikat, penanganan nilai kosong, seleksi fitur berbasis korelasi, dan normalisasi. Model yang diuji meliputi *Logistic Regression* dan *Random Forest* untuk *machine learning*, serta *Sequential* dan *TabNet* untuk *deep learning*. Evaluasi performa dilakukan menggunakan metrik akurasi, presisi, *recall,* dan *F1-score*. Interpretasi model dilakukan dengan metode *Explainable Artificial Intelligence* (XAI) menggunakan SHAP untuk analisis global dan LIME untuk analisis lokal. Hasil penelitian menunjukkan bahwa penerapan SMOTE secara signifikan meningkatkan kinerja model, terutama dalam mendeteksi kelas dengan label Stroke, dengan *Random Forest* pasca-SMOTE mencapai akurasi 97% dan keseimbangan presisi–recall yang optimal. Temuan ini mendukung hipotesis bahwa teknik *balancing data* mampu meningkatkan deteksi kasus berisiko rendah, serta memberikan kontribusi praktis bagi pengembangan sistem skrining kesehatan berbasis kecerdasan buatan yang transparan dan andal.**

**Kata kunci: SMOTE, *machine learning, deep learning*, stroke, XAI**

1. LATAR BELAKANG

Stroke merupakan salah satu penyebab utama kematian dan kecacatan di seluruh dunia. Menurut World Stroke Organization [1], setiap tahunnya lebih dari 12 juta orang mengalami stroke pertama, dan sekitar 6,5 juta di antaranya meninggal akibat kondisi ini. Di Indonesia, prevalensi stroke terus meningkat seiring dengan perubahan gaya hidup masyarakat, peningkatan usia harapan hidup, dan tingginya prevalensi faktor risiko seperti hipertensi, diabetes, obesitas, serta perilaku merokok [2]. Kondisi ini tidak hanya meningkatkan beban perawatan medis, tetapi juga berimplikasi pada menurunnya kualitas hidup penderita serta meningkatnya beban ekonomi bagi keluarga dan negara. Penanganan stroke yang efektif memerlukan strategi komprehensif yang mencakup pencegahan, deteksi dini, diagnosis cepat, dan pengobatan tepat sasaran. Oleh karena itu, diperlukan pendekatan inovatif yang mampu memanfaatkan teknologi modern untuk memperbaiki kualitas prediksi dan deteksi stroke sejak tahap awal.

Teknologi *machine learning* (ML) telah mengalami perkembangan pesat dalam bidang kesehatan, termasuk dalam diagnosis dan prediksi risiko stroke. ML memiliki kemampuan untuk mengolah data medis yang kompleks dan menemukan pola tersembunyi yang sulit dideteksi secara manual oleh tenaga medis. Garg et al. [3] memanfaatkan algoritma *Random Forest* dan *Logistic Regression* untuk memprediksi risiko stroke menggunakan data klinis, dan hasilnya menunjukkan kemampuan generalisasi yang baik pada dataset seimbang. Penelitian lain oleh Prasetyawati et al. [4] membandingkan performa *Support Vector Machine* (SVM) dan *k-Nearest Neighbor* (k-NN), di mana SVM terbukti lebih unggul dalam menangani data multidimensi. Meski demikian, metode ML klasik sering menghadapi tantangan pada *imbalanced dataset*, yaitu kondisi ketika jumlah kasus stroke (kelas minoritas) jauh lebih sedikit dibandingkan dengan non-stroke (kelas mayoritas). Ketidakseimbangan ini menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga sensitivitas terhadap kasus stroke menjadi rendah dan berisiko menghasilkan prediksi yang tidak akurat bagi pasien yang sebenarnya memerlukan perhatian medis.

Selain metode ML klasik, *deep learning* (DL) telah menjadi salah satu teknologi yang sangat menjanjikan, khususnya dalam analisis citra medis. DL memiliki keunggulan dalam mengekstraksi fitur kompleks secara otomatis tanpa memerlukan proses *feature engineering* manual. Dai et al. [5] memanfaatkan arsitektur *Convolutional Neural Networks* (CNN) untuk klasifikasi citra CT pasien stroke dan memperoleh akurasi lebih tinggi dibandingkan metode tradisional. Chen et al. [6] menggunakan ResNet pada data MRI difusi untuk mengidentifikasi area otak yang terdampak stroke dengan presisi lebih tinggi. Walaupun demikian, kinerja DL sangat bergantung pada ketersediaan dataset besar dan seimbang, sedangkan dalam kasus stroke, dataset seperti itu relatif sulit diperoleh. Kondisi ini menegaskan pentingnya penerapan teknik penyeimbangan data seperti *Synthetic Minority Over-sampling Technique* (SMOTE) [7] untuk mengurangi bias terhadap kelas mayoritas dan meningkatkan performa model pada deteksi kasus stroke. Penggunaan SMOTE memungkinkan replikasi data sintetis pada kelas minoritas sehingga distribusi data menjadi lebih proporsional.

Tantangan lain yang sangat penting dalam penerapan teknologi AI di bidang medis adalah masalah interpretabilitas model. Banyak algoritma DL dan sebagian algoritma ML canggih beroperasi sebagai *black box*, artinya proses pengambilan keputusan model sulit dipahami oleh pengguna akhir, dalam hal ini tenaga medis [8]. Hal ini menimbulkan keraguan dan resistensi dalam adopsi teknologi AI untuk pengambilan keputusan klinis, terutama pada kasus kritis seperti stroke. Dalam konteks medis, tenaga kesehatan memerlukan penjelasan yang jelas dan terperinci mengenai alasan di balik prediksi atau diagnosis yang dihasilkan oleh model. Pendekatan *Explainable AI* (XAI) hadir untuk menjawab tantangan ini dengan menyediakan metode interpretasi seperti SHAP dan LIME yang dapat memberikan justifikasi terhadap hasil prediksi [8]. Selain itu, Ghosh et al. [9] menerapkan metode Grad-CAM pada model CNN untuk memvisualisasikan area otak yang menjadi fokus model dalam mendeteksi stroke. Penerapan XAI diharapkan dapat meningkatkan kepercayaan, transparansi, dan akuntabilitas sistem deteksi berbasis AI dalam praktik klinis.

Berdasarkan kajian literatur yang ada, dapat diidentifikasi tiga celah penelitian utama dalam penerapan teknologi AI untuk deteksi stroke. Pertama, performa ML klasik masih terbatas ketika berhadapan dengan dataset stroke yang tidak seimbang, yang mengakibatkan rendahnya sensitivitas model terhadap kelas minoritas. Kedua, DL meskipun unggul pada analisis citra medis, masih sangat bergantung pada dataset besar dan seimbang, sehingga memerlukan strategi penyeimbangan data seperti SMOTE untuk mengoptimalkan kinerjanya. Ketiga, integrasi XAI dalam sistem deteksi stroke masih rendah, padahal interpretabilitas sangat dibutuhkan untuk meningkatkan kepercayaan dan adopsi teknologi di lingkungan medis. Kesenjangan ini menunjukkan bahwa penelitian yang menggabungkan ML klasik dan DL, memanfaatkan teknik penyeimbangan data, dan mengintegrasikan XAI merupakan hal yang mendesak untuk dilakukan. Dengan demikian, penelitian ini diarahkan untuk memberikan solusi komprehensif terhadap ketiga permasalahan tersebut secara bersamaan. Upaya ini diharapkan mampu mendorong penerapan AI yang lebih efektif dan dapat dipercaya di bidang medis, khususnya pada kasus stroke.

Penelitian ini bertujuan untuk mengevaluasi kinerja algoritma ML klasik dan DL dalam klasifikasi penyakit stroke pada dataset yang tidak seimbang, dengan menerapkan SMOTE untuk meningkatkan sensitivitas terhadap kelas minoritas, serta mengintegrasikan metode XAI (SHAP, LIME, dan Grad-CAM) guna meningkatkan transparansi prediksi. Kontribusi utama penelitian ini meliputi tiga hal. Pertama, evaluasi komparatif kinerja algoritma ML klasik dan DL pada data stroke yang tidak seimbang, sehingga diperoleh pemahaman yang komprehensif terkait keunggulan dan keterbatasan masing-masing pendekatan. Kedua, implementasi SMOTE sebagai strategi penyeimbangan data yang efektif untuk mengatasi bias terhadap kelas mayoritas dan meningkatkan deteksi kasus stroke. Ketiga, integrasi XAI ke dalam sistem deteksi stroke untuk memberikan interpretasi prediksi yang mudah dipahami oleh tenaga medis, sehingga hasil model tidak hanya akurat tetapi juga transparan. Dengan pendekatan ini, diharapkan sistem deteksi stroke yang dihasilkan memiliki performa tinggi, mudah diadopsi, dan mampu berkontribusi nyata dalam praktik klinis. Pada akhirnya, penelitian ini diharapkan dapat menjadi pijakan bagi pengembangan teknologi AI yang lebih adaptif, transparan, dan bermanfaat di dunia kesehatan.

1. METHOD

Penelitian ini memanfaatkan data sekunder yang diperoleh dari *Stroke Prediction Dataset* yang tersedia di platform Kaggle, yang memuat berbagai atribut demografis dan klinis yang relevan untuk tujuan prediksi risiko stroke. Metode penelitian ini dirancang untuk mengembangkan sistem deteksi penyakit stroke menggunakan kombinasi *machine learning* (ML), *deep learning* (DL), *feature engineering*, teknik penyeimbangan data, serta metode *Explainable Artificial Intelligence* (XAI). Alur metode terdiri atas tujuh tahapan utama sesuai pada Fig. 1, yaitu: akuisisi data, pra-pemrosesan, *feature engineering*, pemodelan, evaluasi, XAI, dan analisis serta interpretasi hasil.

A diagram of a process

AI-generated content may be incorrect.

Fig. 1 Metode penelitian

Tahap pertama adalah akuisisi data, yang menggunakan *Stroke Prediction Dataset* dari platform Kaggle. Dataset ini mencakup variabel demografis, klinis, dan riwayat kesehatan pasien, serta label biner yang menunjukkan apakah pasien pernah mengalami stroke. Pemilihan dataset ini didasarkan pada kelengkapan atribut yang relevan untuk analisis prediktif. Data yang tersedia memberikan peluang untuk menguji berbagai model *machine learning* (ML) dan *deep learning* (DL) dalam mendeteksi risiko stroke. Selain itu, dataset ini telah digunakan pada berbagai penelitian sebelumnya, sehingga memungkinkan perbandingan hasil secara objektif [4], [10].

Tahap pra-pemrosesan dilakukan untuk meningkatkan kualitas data sebelum masuk ke tahap pemodelan. Langkah yang dilakukan meliputi penghapusan data duplikat (*drop duplicates*), penanganan nilai kosong (*missing values*) melalui metode imputasi [11], dan seleksi fitur menggunakan *correlation matrix* [12]. Seleksi ini hanya mempertahankan fitur yang memiliki korelasi signifikan terhadap variabel target. Pada tahap ini belum dilakukan penyeimbangan data, karena teknik SMOTE akan digunakan terpisah sebagai pembanding performa model. Pendekatan ini memudahkan analisis pengaruh SMOTE terhadap hasil akhir model [7], [13], [14].

Tahap berikutnya adalah *feature engineering* yang dilakukan menggunakan *Principal Component Analysis* (PCA). PCA digunakan untuk mereduksi dimensi data, sehingga mengurangi risiko *overfitting* dan mempercepat proses pelatihan model [15]. Dengan memproyeksikan data ke komponen utama, informasi yang terkandung tetap terjaga meskipun jumlah variabel berkurang. Metode ini sangat bermanfaat untuk data dengan variabel yang saling berkorelasi tinggi. Penerapan PCA juga membantu meningkatkan generalisasi model [15], [16].

Tahap pemodelan melibatkan dua pendekatan utama, yaitu ML klasik dan DL. Model ML klasik yang digunakan meliputi *Random Forest*, *Logistic Regression*, *Support Vector Machine* (SVM), dan *k-Nearest Neighbor* (k-NN) [3], [4]. Model DL yang digunakan adalah *Convolutional Neural Network* (CNN) yang diadaptasi untuk data tabular [5], [6]. Setiap model diuji dalam dua skenario: tanpa SMOTE dan dengan SMOTE. Teknik SMOTE digunakan untuk mengatasi masalah ketidakseimbangan kelas dengan melakukan sintesis data minoritas [Chawla et al., 2002] [17].

Evaluasi model dilakukan dengan mengukur akurasi, presisi, recall, F1-score, dan AUC-ROC [18]. Metrik ini dipilih karena memberikan gambaran menyeluruh mengenai performa model dalam mendeteksi kelas minoritas. Presisi dan recall berfokus pada keseimbangan antara prediksi positif yang benar dan kemampuan mendeteksi kasus positif. F1-score memberikan rata-rata harmonis dari kedua metrik tersebut, sementara AUC-ROC mengukur kemampuan model dalam membedakan antara kelas positif dan negatif. Pendekatan evaluasi ini memastikan model terbaik dipilih secara objektif [3], [5].

Tahap *explainable AI* (XAI) dilakukan untuk memberikan interpretasi terhadap prediksi yang dihasilkan model. Dua metode yang digunakan adalah SHAP (*SHapley Additive exPlanations*) dan LIME (*Local Interpretable Model-agnostic Explanations*) [8], [19]. SHAP digunakan untuk mengukur kontribusi setiap fitur terhadap hasil prediksi model secara global [20]. LIME digunakan untuk menjelaskan prediksi individu secara lokal, sehingga memudahkan interpretasi kasus spesifik [20], [21]. Penggunaan XAI sangat penting untuk meningkatkan kepercayaan tenaga medis terhadap hasil sistem prediksi [8], [19].

Tahap terakhir adalah analisis dan interpretasi hasil. Pada tahap ini, hasil evaluasi model dibandingkan antara skenario tanpa SMOTE dan dengan SMOTE. Faktor-faktor penting yang mempengaruhi prediksi dianalisis berdasarkan output SHAP dan LIME. Model terbaik dipilih berdasarkan kombinasi kinerja dan interpretabilitas. Rekomendasi model yang dihasilkan diharapkan dapat diimplementasikan pada sistem pendukung keputusan medis untuk deteksi dini risiko stroke [10], [13].

1. RESULT AND DISCUSSION

Penelitian ini menggunakan dua pendekatan utama, yaitu *machine learning* dan *deep learning*, untuk membandingkan kinerja deteksi kasus stroke pada dataset dengan kondisi seimbang maupun tidak seimbang. Pada pendekatan *machine learning*, algoritma yang digunakan adalah *Logistic Regression* dan *Random Forest*, yang dievaluasi pada kelas *No Stroke* dan *Stroke*. Sementara itu, pada pendekatan *deep learning*, digunakan dua arsitektur yang berbeda, yaitu model *Sequential* dan *TabNet Classifier*. Kedua pendekatan ini diuji dalam dua skenario, yaitu sebelum dan sesudah penerapan teknik SMOTE, untuk mengukur dampak penyeimbangan data terhadap kemampuan deteksi kelas minoritas.

1. *Machine Learning*

Penelitian ini menerapkan dua algoritma *machine learning*, yaitu *Logistic Regression* dan *Random Forest*. Kinerja kedua model dievaluasi menggunakan metrik *precision*, *recall*, F1-score, serta akurasi untuk masing-masing kelas *No Stroke* dan *Stroke*. Ringkasan hasil evaluasi disajikan dalam tabel berikut untuk mempermudah analisis dan perbandingan performa secara komprehensif.

TABLE 1

Hasil model *machine learning*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **SMOTE** | **Kelas** | ***Precision*** | ***Recall*** | ***F1-Score*** | ***Support*** | **Akurasi** |
| *Logistic Regression* | Tidak | *No Stroke* | 0.99 | 0.73 | 0.84 | 245 | 0.73 |
| *Stroke* | 0.07 | 0.83 | 0.13 | 6 |
| *Random Forest* | Tidak | *No Stroke* | 0.98 | 1.00 | 0.99 | 246 | 0.98 |
| *Stroke* | 0.00 | 0.00 | 0.00 | 5 |
| *Logistic Regression* | Ya | *No Stroke* | 0.91 | 0.87 | 0.89 | 236 | 0.89 |
| *Stroke* | 0.88 | 0.91 | 0.90 | 244 |
| *Random Forest* | Ya | *No Stroke* | 0.96 | 0.98 | 0.97 | 233 | 0.97 |
| *Stroke* | 0.98 | 0.96 | 0.97 | 247 |

Hasil evaluasi model *machine learning* tanpa penerapan SMOTE menunjukkan adanya ketimpangan performa yang cukup signifikan antara kelas mayoritas (*No Stroke*) dan kelas minoritas (*Stroke*). Model *Logistic Regression* memperoleh akurasi sebesar 73%, dengan nilai *precision* untuk kelas *Stroke* hanya sebesar 0,07, sementara nilai *recall*-nya cukup tinggi yaitu 0,83. Kondisi ini menunjukkan bahwa meskipun model relatif sering memprediksi kelas *Stroke*, sebagian besar prediksi tersebut ternyata keliru (*false positive*), sehingga menurunkan nilai F1-Score secara signifikan. Sementara itu, model Random Forest menunjukkan akurasi yang sangat tinggi, yaitu 98%, namun gagal mendeteksi kelas *Stroke* sama sekali dengan nilai *recall* = 0,00. Hal ini mengindikasikan bahwa model tersebut sangat bias terhadap kelas mayoritas, hanya memprediksi *No Stroke* dan mengabaikan deteksi kasus *Stroke*. Fenomena ini mendukung hipotesis awal penelitian bahwa ketidakseimbangan data akan berdampak langsung pada kemampuan model dalam mendeteksi kelas minoritas.

Hasil evaluasi model *machine learning* tanpa penerapan SMOTE menunjukkan adanya ketimpangan performa yang cukup signifikan antara kelas mayoritas (*No Stroke*) dan kelas minoritas (*Stroke*). Model *Logistic Regression* memperoleh akurasi sebesar 73%, dengan nilai *precision* untuk kelas *Stroke* hanya sebesar 0,07, sementara nilai *recall*-nya cukup tinggi yaitu 0,83. Kondisi ini menunjukkan bahwa meskipun model relatif sering memprediksi kelas *Stroke*, sebagian besar prediksi tersebut ternyata keliru (*false positive*), sehingga menurunkan nilai F1-Score secara signifikan. Sementara itu, model Random Forest menunjukkan akurasi yang sangat tinggi, yaitu 98%, namun gagal mendeteksi kelas *Stroke* sama sekali dengan nilai *recall* = 0,00. Hal ini mengindikasikan bahwa model tersebut sangat bias terhadap kelas mayoritas, hanya memprediksi *No Stroke* dan mengabaikan deteksi kasus *Stroke*. Fenomena ini mendukung hipotesis awal penelitian bahwa ketidakseimbangan data akan berdampak langsung pada kemampuan model dalam mendeteksi kelas minoritas.

Setelah penerapan teknik *oversampling* menggunakan SMOTE, performa kedua model mengalami peningkatan yang signifikan. *Logistic Regression* yang sebelumnya memiliki kesenjangan performa antarkelas, setelah SMOTE mencapai akurasi sebesar 89% dengan *precision* dan *recall* yang seimbang pada kedua kelas. Hal ini menandakan bahwa distribusi data yang lebih seimbang membuat model dapat mengenali pola dari kelas minoritas dengan lebih baik. Random Forest pasca-SMOTE bahkan mencatatkan akurasi 97%, dengan *precision* dan *recall* yang keduanya bernilai 0,97 untuk kelas *Stroke*, menunjukkan deteksi yang nyaris sempurna. Peningkatan performa ini membuktikan bahwa SMOTE efektif dalam mengurangi bias model terhadap kelas mayoritas. Dengan kata lain, teknik penyeimbangan data menjadi salah satu faktor kunci yang mempengaruhi performa model dalam skenario klasifikasi medis dengan distribusi data yang timpang.

Secara keseluruhan, hasil ini memberikan gambaran bahwa tanpa penanganan khusus, model cenderung mengabaikan kelas minoritas yang dalam konteks medis justru menjadi kelompok penting untuk diidentifikasi. Penerapan SMOTE mampu mengubah pola pembelajaran model, sehingga deteksi kelas minoritas meningkat drastis tanpa mengorbankan performa pada kelas mayoritas. Hal ini relevan dalam kasus prediksi penyakit seperti stroke, di mana identifikasi kasus positif lebih krusial dibanding sekadar mempertahankan akurasi tinggi. Temuan ini juga memperkuat pentingnya tahap pra-pemrosesan dan penyeimbangan data sebelum pelatihan model *machine learning*. Dengan distribusi data yang seimbang, baik algoritme linear seperti *Logistic Regression* maupun algoritme berbasis pohon seperti Random Forest mampu memberikan performa yang stabil. Keberhasilan ini membuka peluang penerapan pendekatan serupa pada berbagai dataset medis lain yang memiliki karakteristik *imbalanced.*

1. *Deep Learning*

Pada pendekatan *deep learning*, penelitian ini menggunakan dua arsitektur, yaitu model *Sequential* dan *TabNet Classifier*. Kinerja kedua model tersebut diukur menggunakan metrik *precision*, *recall*, *F1-score*, serta akurasi pada kelas *No Stroke* dan *Stroke*. Hasil pengujian dirangkum dalam tabel berikut guna memudahkan analisis serta perbandingan performa secara menyeluruh pada berbagai kondisi data.

TABLE 2

Hasil model *deep learning*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **SMOTE** | **Kelas** | **Precision** | **Recall** | **F1-Score** | **Support** | **Akurasi** |
| *Sequential* | Tidak | *No Stroke* | 0.96 | 1.00 | 0.98 | 2965 | 0.96 |
| *Stroke* | 0.00 | 0.00 | 0.00 | 128 |
| *Sequential* | Ya | *No Stroke* | 0.92 | 0.90 | 0.91 | 190 | 0.91 |
| *Stroke* | 0.90 | 0.92 | 0.91 | 186 |
| *TabNet Classifier* | Tidak | *No Stroke* | 0.97 | 0.81 | 0.88 | 4089 | 0.80 |
| *Stroke* | 0.08 | 0.38 | 0.14 | 182 |
| *TabNet Classifier* | Ya | *No Stroke* | 0.97 | 0.94 | 0.95 | 94 | 0.95 |
| *Stroke* | 0.94 | 0.97 | 0.95 | 94 |

Hasil evaluasi model *deep learning* tanpa penerapan SMOTE menunjukkan pola permasalahan yang serupa dengan temuan pada model *machine learning*, yakni bias kuat terhadap kelas dengan label *No Stroke*. Model Sequential tanpa SMOTE mencatat akurasi 96%, namun gagal mendeteksi kasus *Stroke* sama sekali dengan nilai *precision* dan *recall* untuk kelas minoritas sebesar 0,00. Hal ini menunjukkan bahwa meskipun akurasi keseluruhan terlihat tinggi, kemampuan deteksi terhadap kasus yang paling kritis dalam konteks medis justru nihil. Model *TabNet* *Classifier* tanpa SMOTE menunjukkan sedikit perbaikan dengan nilai *recall* untuk kelas *Stroke* sebesar 0,38, tetapi masih jauh dari ideal. Nilai *precision* untuk kelas *Stroke* pada *TabNet* hanya 0,08, yang mengindikasikan banyak prediksi positif palsu (*false positive*). Kondisi ini kembali menegaskan bahwa ketidakseimbangan data menyebabkan *deep learning* cenderung mengoptimalkan kinerja pada kelas mayoritas, mengorbankan kemampuan deteksi pada kelas minoritas.

Penerapan SMOTE memberikan dampak yang signifikan pada kinerja model *deep learning*. Sequential setelah SMOTE mengalami peningkatan akurasi menjadi 91%, dengan nilai *precision* dan *recall* yang seimbang (0,90–0,92) pada kedua kelas. Hal ini menunjukkan bahwa penyeimbangan data membantu model mengenali pola-pola penting pada kelas minoritas tanpa mengurangi kemampuan mendeteksi kelas mayoritas. Model *TabNet Classifier* pasca-SMOTE menunjukkan peningkatan performa yang sangat signifikan, mencapai akurasi 95% dengan *precision* dan *recall* sama-sama 0,95 untuk kelas *Stroke*. Performa yang seimbang ini menghasilkan nilai *F1-Score* yang tinggi, mengindikasikan bahwa model mampu mempertahankan tingkat prediksi positif yang benar sambil meminimalkan kesalahan klasifikasi. Peningkatan ini tidak hanya relevan secara statistik, tetapi juga sangat penting secara klinis karena membantu memastikan bahwa pasien berisiko stroke dapat terdeteksi dengan akurat.

Jika dibandingkan, dampak SMOTE pada *deep learning* bahkan terlihat lebih konsisten dibandingkan pada *machine learning*, karena kedua model Sequential dan *TabNet* sama-sama menunjukkan keseimbangan performa yang hampir sempurna pasca penyeimbangan data. Perbedaan utama terletak pada kemampuan *TabNet* pasca-SMOTE yang mampu mempertahankan akurasi tinggi dengan dataset yang relatif lebih kecil dibanding Sequential. Hal ini mengindikasikan bahwa *TabNet* dapat menjadi pilihan yang lebih efisien ketika sumber daya komputasi terbatas, tanpa mengorbankan kualitas deteksi. Temuan ini kembali menegaskan bahwa dalam konteks data medis yang imbalanced, tahap pra-pemrosesan seperti SMOTE menjadi langkah wajib untuk mengoptimalkan performa model. Lebih jauh, pendekatan ini dapat diadaptasi pada skenario diagnosis penyakit lainnya yang memiliki distribusi kelas timpang. Dengan demikian, hasil penelitian ini membuka ruang untuk eksplorasi model *deep learning* berbasis data seimbang pada berbagai domain medis yang memerlukan deteksi dini.

1. *Xai SHAP*

Hasil analisis menggunakan SHAP (*SHapley Additive exPlanations*) yang ditampilkan pada Fig. 2 menunjukkan bahwa terdapat tiga fitur utama yang memberikan kontribusi terbesar terhadap prediksi risiko stroke, yaitu usia, status merokok, dan jenis pekerjaan. Fitur usia memberikan hubungan positif yang sangat kuat terhadap peningkatan probabilitas stroke, sehingga individu dengan usia lanjut cenderung memiliki risiko yang lebih tinggi. Sebaliknya, status merokok kategori "*never smoked*" terlihat berkontribusi menurunkan risiko stroke secara signifikan. Jenis pekerjaan juga memegang peran penting, kemungkinan besar dipengaruhi oleh faktor seperti tingkat stres, pola aktivitas fisik, atau paparan risiko lingkungan tertentu. Analisis ini sejalan dengan temuan medis sebelumnya yang menyatakan bahwa faktor demografi dan gaya hidup merupakan prediktor utama stroke.

A graph with red and blue lines

AI-generated content may be incorrect.

Fig. 2 *Summary plot* hasil SHAP

Visualisasi *waterfall plot* SHAP yang diperlihatkan pada Fig. 3 memberikan pandangan detail mengenai kontribusi tiap fitur dalam sebuah kasus prediksi stroke. Pada contoh kasus tersebut, faktor-faktor seperti tinggal di wilayah urban, jenis kelamin laki-laki, dan status "*never smoked*" terlihat menurunkan probabilitas stroke. Namun, usia tinggi dan jenis pekerjaan tertentu justru memberikan kontribusi positif yang signifikan dalam mendorong hasil prediksi ke arah positif (stroke). Informasi ini penting karena dapat mengidentifikasi faktor risiko yang dapat dimodifikasi (*modifiable risk factors*), seperti perubahan gaya hidup atau intervensi lingkungan kerja. Dengan adanya *waterfall plot*, proses pengambilan keputusan menjadi lebih transparan bagi tenaga medis maupun pasien.

A screenshot of a computer

AI-generated content may be incorrect.

Fig. 3 *Force plot* hasil SHAP

Selain itu, Fig. 4 menampilkan visualisasi kontribusi rata-rata fitur yang dihasilkan dari SHAP, yang kemudian dibandingkan dengan koefisien pada model *Logistic Regression*. Hasil perbandingan ini menunjukkan konsistensi yang kuat, di mana usia dan variabel gaya hidup seperti kebiasaan merokok tetap menjadi faktor dominan dalam kedua analisis. Konsistensi tersebut memperkuat validitas temuan dan mengurangi potensi bias yang mungkin muncul dari penggunaan model tunggal. Temuan ini memiliki implikasi praktis bagi strategi pencegahan stroke, khususnya pada pengendalian kebiasaan merokok dan perawatan kesehatan pada usia lanjut. Penyelarasan hasil dari SHAP dan *Logistic Regression* juga meningkatkan kepercayaan terhadap sistem prediksi yang dikembangkan.

A graph with blue bars

AI-generated content may be incorrect.

Fig. 4 *Bar plot* *logistic regression coefficients*

1. *XAI LIME*

Pendekatan *Local Interpretable Model-Agnostic Explanations* (LIME) digunakan dalam penelitian ini untuk memberikan penjelasan terperinci terhadap prediksi individu yang dihasilkan oleh model *Random Forest*. Teknik ini berfokus pada interpretasi lokal, sehingga memungkinkan pemahaman yang lebih mendalam mengenai alasan di balik suatu prediksi spesifik. Pada contoh kasus yang dianalisis, probabilitas model untuk memprediksi kelas "*No Stroke*" mencapai 96%, sedangkan probabilitas untuk kelas "Stroke" hanya sebesar 4% pada Fig. 5. Perbedaan probabilitas yang cukup signifikan ini mengindikasikan bahwa model memiliki keyakinan tinggi pada hasil prediksi *No Stroke*. LIME membantu menguraikan faktor-faktor yang memengaruhi probabilitas tersebut, dengan membedakan kontribusi positif dan negatif terhadap hasil akhir.

A screenshot of a computer

AI-generated content may be incorrect.

Fig. 5 *Lime output*

Berdasarkan visualisasi LIME pada Fig. 5, beberapa faktor memberikan kontribusi negatif terhadap kemungkinan prediksi stroke, sehingga memperkuat klasifikasi "No Stroke". Salah satu faktor dominan adalah nilai usia yang telah dinormalisasi ≤ -0,59, yang menunjukkan usia relatif muda dalam konteks dataset ini. Selain itu, status merokok tertentu misalnya *never smoked* atau *formerly smoked* juga berperan menurunkan risiko stroke. Jenis pekerjaan tertentu turut menjadi faktor yang signifikan dalam menekan probabilitas stroke, kemungkinan karena tingkat aktivitas fisik atau paparan risiko lingkungan yang lebih rendah. Kombinasi faktor-faktor ini menyebabkan penurunan probabilitas yang cukup besar pada kelas stroke, sehingga menggeser prediksi model ke arah kelas "No Stroke". Adanya hubungan visual dan naratif yang jelas pada Fig. 5 membantu pembaca memahami bagaimana faktor-faktor ini bekerja secara simultan.

Meskipun mayoritas faktor pada Fig. 5 menunjukkan kontribusi negatif terhadap risiko stroke, terdapat pula variabel yang memberikan kontribusi positif walaupun tidak dominan. Faktor-faktor tersebut meliputi riwayat penyakit jantung, hipertensi, jenis kelamin laki-laki, serta tempat tinggal di wilayah urban. Variabel-variabel ini meningkatkan probabilitas stroke dalam jumlah kecil, namun tetap relevan untuk interpretasi risiko. Keberadaan faktor positif ini menunjukkan bahwa meskipun pasien diprediksi "*No Stroke*", terdapat kondisi tertentu yang perlu mendapat perhatian lebih lanjut. Analisis LIME ini melengkapi temuan global dari SHAP, dengan memberikan gambaran personal yang lebih kontekstual bagi setiap individu. Integrasi hasil dari Fig. 5 dengan visualisasi SHAP memberikan pendekatan yang lebih komprehensif, meningkatkan transparansi serta akuntabilitas sistem prediksi berbasis *machine learning*.

1. Discussion

Hasil penelitian ini menunjukkan bahwa penerapan *Synthetic Minority Oversampling Technique* (SMOTE) secara signifikan meningkatkan kemampuan model, baik pada pendekatan *machine learning* pada model *Logistic Regression* dan *Random Forest* maupun *deep learning* pada model *Sequential dan TabNet Classifier*, dalam mendeteksi kasus stroke. Temuan ini mengonfirmasi hipotesis awal bahwa ketidakseimbangan data merupakan faktor yang menurunkan kinerja model, dan teknik *oversampling* mampu memperbaikinya [17], [22]. Hal ini konsisten dengan teori yang diuraikan pada tinjauan pustaka, di mana penyeimbangan distribusi kelas dapat memperbaiki *recall* kelas minoritas tanpa mengorbankan akurasi keseluruhan. Analisis XAI menggunakan SHAP memberikan wawasan global mengenai peran fitur seperti usia, status merokok, dan jenis pekerjaan, yang sejalan dengan literatur epidemiologi stroke [23]. Sementara itu, LIME memberikan interpretasi lokal yang memperlihatkan variasi kontribusi faktor pada setiap prediksi individual [21]. Kombinasi SMOTE dan XAI ini mendukung pandangan teoritis bahwa akurasi dan interpretabilitas perlu berjalan beriringan untuk memastikan keandalan sistem pendukung keputusan di bidang kesehatan [8].

Dibandingkan dengan penelitian sebelumnya, hasil ini sejalan dengan studi oleh Chawla et al. dan Fernández et al. yang menemukan bahwa SMOTE mampu meningkatkan deteksi kelas minoritas pada domain kesehatan. Namun, tingkat peningkatan pada penelitian ini lebih tinggi, khususnya pada model *Random Forest* pasca-SMOTE, yang mungkin disebabkan oleh kombinasi fitur yang relevan secara klinis dan preprocessing data yang optimal [24]. Berbeda dengan temuan Zhang et al. yang melaporkan bahwa *deep learning* tanpa balancing data sudah cukup akurat, penelitian ini menunjukkan bahwa balancing tetap diperlukan untuk meningkatkan deteksi kasus stroke yang jarang. Perbedaan ini dapat dijelaskan oleh variasi karakteristik dataset, ukuran sampel, serta kompleksitas fitur yang digunakan. Selain itu, integrasi SHAP dan LIME memberikan nilai tambah yang tidak selalu digunakan pada studi sebelumnya, sehingga memberikan pemahaman yang lebih komprehensif terhadap mekanisme prediksi [20], [21]. Faktor metodologis ini menjadi pembeda utama yang mungkin memengaruhi perbedaan hasil antara penelitian ini dan studi terdahulu.

Implikasi praktis dari penelitian ini adalah perlunya penerapan metode balancing data dan interpretabilitas model pada sistem pendukung keputusan medis, khususnya untuk penyakit dengan insidensi rendah seperti stroke [23]. Hasil penelitian dapat digunakan oleh institusi kesehatan untuk mengembangkan algoritma skrining yang lebih sensitif terhadap kasus berisiko tinggi, sekaligus tetap mempertahankan transparansi dalam pengambilan keputusan klinis [8]. Secara teoretis, penelitian ini berkontribusi pada pengembangan pendekatan hibrida yang memadukan *data balancing* dan *explainable AI* sebagai strategi meningkatkan akurasi dan kepercayaan pengguna. Namun, penelitian ini memiliki keterbatasan, seperti penggunaan dataset tunggal dan belum menguji model pada data real-time atau lintas populasi [24]. Untuk penelitian selanjutnya, disarankan melakukan validasi eksternal menggunakan data multi-sumber, mengeksplorasi teknik balancing yang lebih adaptif, serta membandingkan dengan metode interpretabilitas lain seperti *counterfactual explanations*. Dengan demikian, keberlanjutan riset dapat memperkuat aplikasi praktis dan relevansi akademis temuan ini di masa depan.

1. CONCLUSION

Penelitian ini membuktikan bahwa penerapan *Synthetic Minority Oversampling Technique* (SMOTE) secara signifikan meningkatkan kemampuan *machine learning* menggunakan model *Logistic Regression* dan *Random Forest* maupun *deep learning* dengan model *Sequential* dan *TabNet Classifier* dalam mendeteksi kasus stroke. Temuan ini secara langsung mendukung hipotesis awal bahwa ketidakseimbangan data menjadi faktor utama penurunan kinerja model, dan bahwa teknik oversampling mampu memperbaiki performa deteksi kelas minoritas. Analisis *Explainable Artificial Intelligence* (XAI) melalui SHAP dan LIME memberikan wawasan global dan lokal yang konsisten, menunjukkan bahwa fitur usia, status merokok, dan jenis pekerjaan merupakan determinan utama dalam prediksi stroke.

Secara praktis, hasil penelitian ini memiliki implikasi penting bagi pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya pada deteksi dini penyakit dengan prevalensi rendah seperti stroke. Penerapan metode balancing data dan interpretabilitas model dapat meningkatkan sensitivitas deteksi pasien berisiko tinggi sekaligus mempertahankan transparansi dalam proses pengambilan keputusan medis. Implementasi temuan ini dapat memperkuat strategi skrining populasi dan membantu penyusunan kebijakan kesehatan berbasis data yang lebih akurat.

Namun, penelitian ini memiliki keterbatasan, yaitu penggunaan dataset tunggal dan belum dilakukan validasi eksternal pada populasi berbeda. Hal ini dapat membatasi generalisasi temuan pada konteks yang lebih luas. Oleh karena itu, penelitian lanjutan disarankan untuk menggunakan data multi-sumber, menguji teknik balancing yang lebih adaptif, serta mengeksplorasi metode interpretabilitas tambahan guna memperkaya pemahaman mekanisme prediksi. Dengan pengembangan tersebut, diharapkan sistem prediksi stroke yang dihasilkan dapat lebih andal, adaptif, dan aplikatif pada berbagai kondisi populasi di dunia nyata.

REFERENCES

[1] World Stroke Organization, “Impact of Stroke,” 2023.

[2] Centers for Disease Control and P. (CDC), “Stroke Facts,” 2022.

[3] N. Garg, H. Patel, and A. Sharma, “Predictive Modeling for Stroke Diagnosis Using Machine Learning Algorithms,” *Sci Rep*, vol. 14, no. 1, pp. 1–10, 2024, doi: 10.1038/s41598-024-61665-4.

[4] R. D. Prasetyawati, A. A. Nugroho, and R. Kurniawan, “Comparison of Machine Learning and Deep Learning Techniques for Stroke Prediction,” *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 2, pp. 45–53, 2024.

[5] C. Dai, W. Zhou, Y. Liu, and others, “Deep learning for automatic detection of acute ischemic lesions on diffusion-weighted imaging,” *Methods Inf Med*, vol. 62, no. 3, pp. 103–110, 2023, doi: 10.1016/j.zefq.2023.101739.

[6] C. Chen, W. Zhao, J. Li, and others, “Stroke severity prediction using deep learning-based analysis of brain MRI,” *Front Neurosci*, vol. 16, p. 945722, 2022, doi: 10.3389/fnins.2022.945722.

[7] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Artificial Intelligence and Machine Learning for Medical Imaging,” *Radiographics*, vol. 41, no. 2, pp. 581–599, 2021, doi: 10.1148/rg.2021200169.

[8] G. Vilone and L. Longo, “Explainable Artificial Intelligence: A Systematic Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, 2021, doi: 10.3390/e23010018.

[9] T. Ghosh, M. Jahan, and A. Singh, “The efficacy of electroencephalogram neurofeedback training in cognition, anxiety, and depression in alcohol dependence syndrome: A case study,” *Ind Psychiatry J*, vol. 23, no. 2, p. 166, 2014, doi: 10.4103/0972-6748.151705.

[10] R. K. Yadav, S. Natarajan, and S. Radhakrishnan, “Machine Learning and Deep Learning Algorithms in Stroke Medicine,” *Cureus*, vol. 16, no. 2, p. e38292, 2024, doi: 10.7759/cureus.38292.

[11] E. Rahm and H. H. Do, “Data Cleaning: Problems and Current Approaches,” 2000. [Online]. Available: https://www.researchgate.net/publication/220282831

[12] P. Schober and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” *Anesth Analg*, vol. 126, no. 5, pp. 1763–1768, May 2018, doi: 10.1213/ANE.0000000000002864.

[13] A. Hosny, C. Parmar, J. Quackenbush, L. H. Schwartz, and H. J. Aerts, “The Evolution of Artificial Intelligence in Medical Imaging,” *Nat Biomed Eng*, vol. 7, pp. 995–1010, 2023, doi: 10.1038/s41551-023-01031-6.

[14] M. Sulistiyono, Y. Pristyanto, S. Adi, and G. Gumelar, “SISTEMASI: Jurnal Sistem Informasi Implementasi Algoritma Synthetic Minority Over-Sampling Technique untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi,” 2021. [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[15] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” Apr. 13, 2016, *Royal Society of London*. doi: 10.1098/rsta.2015.0202.

[16] R. Huang, L. Wang, J. Xu, and others, “Comparative Analysis of Machine Learning and Deep Learning Models in Clinical Stroke Prediction,” *Sci Rep*, vol. 14, no. 1, p. 231, 2024, doi: 10.1038/s41598-024-82931-5.

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

[18] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets,” *PLoS One*, vol. 10, no. 3, Mar. 2015, doi: 10.1371/journal.pone.0118432.

[19] E. Tjoa and C. Guan, “Transparency of Deep Neural Networks for Medical Image Analysis: A Review of Interpretability Methods,” 2021.

[20] S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” Nov. 2017, [Online]. Available: http://arxiv.org/abs/1705.07874

[21] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[22] A. Fernández, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. 2018. doi: 10.1007/978-3-319-98074-4.

[23] V. L. Feigin *et al.*, “Global, regional, and national burden of stroke and its risk factors, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021,” *Lancet Neurol*, vol. 23, no. 10, pp. 973–1003, Oct. 2024, doi: 10.1016/S1474-4422(24)00369-7.

[24] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” 2015, *Elsevier B.V.* doi: 10.1016/j.csbj.2014.11.005.