JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER

p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

Analisis Prediksi Risiko Stroke Menggunakan Metode SGD

Febriani Yolanda Tessalonika[1], Laudya Meitaneia Sianturi[2], Esthi Nurani Sri Handayani[3], Sischa Wahyuning Tyas[4], Anggriani Puspita Sari[5,*]

[1,2,3,4]Ilmu Komputer, Sains Data, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
[5]Ilmu Komputer, Informatika, Universitas Pembangunan Nasional Veteran Jawa Timur, Surabaya, Indonesia
Email: [1]febrianiyolanda63@gmail.com, [2]siantrilaudya08@gmail.com, [3]esthinu@gmal.com, [4]sischa_wahyuning.sada@upnjatim.ac.id, [5,*]anggraini.puspita.if@upnjatim.ac.id.
(*Email Corresponding Author: anggraini.puspita.if@upnjatim.ac.id)
Received: 9 Januari 2026 | Revision: 15 Januari 2026 | Accepted: 16 Januari 2026

Abstrak

Stroke merupakan penyebab kematian dan kecacatan tertinggi di dunia dengan lebih dari 15 juta kasus setiap tahun. Deteksi dini risiko stroke sangat penting untuk menurunkan angka mortalitas melalui intervensi preventif yang tepat waktu. Penelitian ini bertujuan untuk menganalisis performa *Stochastic Gradient Descent* (SGD) Regresi Logistik dalam prediksi risiko stroke menggunakan dataset Prediksi Stroke v2 2024 yang terdiri dari 35.000 *records* dengan 17 fitur klinis. Dataset dibagi menjadi 80% *data training* (28.000 *samples*) dan 20% *data testing* (7.000 *samples*) dengan distribusi kelas seimbang. Optimasi hyperparameter SGD dilakukan menggunakan *Grid Search* dengan 216 kombinasi parameter dan 4-*fold cross validation* untuk mendapatkan konfigurasi optimal. Hasil penelitian menunjukkan bahwa SGD Regresi Logistik dengan hyperparameter tuning mencapai performa *excellent* dengan *accuracy* 97,59% dan ROC-AUC 0,9984, menunjukkan sensitivitas sangat tinggi dalam mendeteksi pasien berisiko dengan *recall* 98,53% pada kelas positif. Performa ini sebanding dengan *Baseline* Regresi Logistik yang mencapai *accuracy* 97,74% dengan ROC-AUC identik 0,9984, mengkonfirmasi efektivitas metode SGD untuk klasifikasi risiko stroke. Analisis feature importance mengidentifikasi Age (koefisien 4,868), High Blood Pressure (2,360), dan Chest Pain (1,977) sebagai prediktor terkuat. Model SGD menunjukkan *false negative rate* rendah hanya 1,47%, temuan ini mengkonfirmasi bahwa SGD regresi logistik dengan optimasi hyperparameter sistematis memberikan performa yang superior dalam sensitivitas deteksi pasien beresiko, menjadikannya sangat potensial untuk diimplementasikan sebagai *screening tool* awal risiko stroke dengan keunggulan *scalability* untuk dataset yang lebih besar dan kemampuan *online learning* untuk *data streaming* yang tidak dimiliki baseline model.

Kata Kunci: Stroke, Faktor Klinis, *Stochastic Gradient Descent*, Prediksi Risiko, Regresi Logistik.

Abstract

Stroke is the leading cause of death and disability worldwide, with more than 15 million cases each year. Early detection of stroke risk is crucial to reducing mortality rates through timely preventive interventions. This study aims to analyze the performance of Stochastic Gradient Descent (SGD) Logistic Regression in predicting stroke risk using the Stroke Prediction v2 2024 dataset, which consists of 35,000 records with 17 clinical features. The dataset was divided into 80% training data (28,000 samples) and 20% testing data (7,000 samples) with balanced class distribution. SGD hyperparameter optimization was performed using Grid Search with 216 parameter combinations and 4-fold cross validation to obtain the optimal configuration. The results show that SGD Logistic Regression with hyperparameter tuning achieved excellent performance with an accuracy of 97.59% and ROC-AUC of 0.9984, indicating very high sensitivity in detecting at-risk patients with a recall of 98.53% in the positive class. This performance is comparable to Baseline Logistic Regression, which achieved an accuracy of 97.74% with an identical ROC-AUC of 0.9984, confirming the effectiveness of the SGD method for stroke risk classification. Feature importance analysis identified Age (coefficient 4.868), High Blood Pressure (2.360), and Chest Pain (1.977) as the strongest predictors. The SGD model shows a low false negative rate of only 1.47%. This finding confirms that logistic regression SGD with systematic hyperparameter optimization provides superior performance in detecting high-risk patients, making it highly potential for implementation as an initial stroke risk screening tool with the advantages of scalability for larger datasets and online learning capabilities for streaming data, which the baseline model does not have.

Keywords: *Stroke, Clinical Factors, Stochastic Gradient Descent, Risk Prediction, Logistic Regression.*

1. PENDAHULUAN

Stroke merupakan penyakit pertama yang diprediksi sebagai penyebab utama kecacatan permanen dalam angka kematian yang dipicu oleh faktor usia, pola hidup dan aktivitas fisik, di antara gangguan tidak menular, stroke tetap menjadi penyebab kematian kedua dan penyebab kematian dan kecacatan gabungan ketiga [1]. Studi epidemiologi menunjukkan insidensi stroke, hipertensi, diabetes melitus dan kadar gula yang tidak terkontrol menjadi awal gejala stroke yang sering dianggap sebagai *"silent killer"* [1], [2]. Rusaknya dinding pembuluh darah di otak pada pasien hipertensi menjadi risiko tertinggi yang dipengaruhi oleh tekanan darah sistolik dengan rasio hiperglikemia stress [2], [3]. Orang dewasa muda mengalami insidensi stroke iskemik mengidentifikasi beberapa variabel klinis penting seperti HbA1c, rasio kolesterol, dan durasi diabetes [4]. Sesuai dengan identifikasi faktor stroke, usia dan kelamin menjadi penyebab stroke pada proses perubahan struktur yang semakin menua atau disfungsi pembuluh darah dan jenis kelamin menjelaskan populasi laki-laki memiliki risiko stroke lebih tinggi karena pola gaya hidup yang disebabkan seseorang yang mengkonsumsi rokok akan

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

mengalami komplikasi paru-paru yang akan berpengaruh kepada turunnya aktivitas fisik karena tekanan darah buruk. Laki-laki cenderung mengalami berat badan atau obesitas sentral dan sistem vaskular yang akan mengakibatkan mudah kelelahan sehingga memiliki risiko stroke iskemik maupun hemoragik [5].

Risiko stroke pada faktor seperti hipertensi menjadikan tekanan darah yang tidak terkontrol mempercepat kerusakan pada vaskular yang memungkinkan terjadinya gangguan aliran darah ke otak dan jantung yang akan mengalami sesak napas, batuk berkepanjangan serta henti napas. Gejala ini memberikan sinyal pada peningkatan beban kerja jantung menyebabkan stres hemodinamik dan ketidakseimbangan aktivitas miokard. Orang dengan *atrial fibrillation* menunjukkan *perfusi miokard* yang kurang optimal secara signifikan terhadap kontrol, yang selaras dengan mekanisme iskemia dan potensi gejala nyeri dada yang menurunkan pemompaan darah [6]. Ketika kebutuhan jantung meningkat akan memunculkan nyeri dada sehingga mengalami ketidaknyamanan pada dada. Gangguan aliran darah akibat penyempitan arteri secara klinis akan muncul kulit yang pucat dan dingin, sensasi kebas atau kesemutan, serta kelemahan otot akibat suplai oksigen yang tidak mencukupi di jaringan distal sehingga terjadi stroke iskemik [7].

Sekitar 13,17% kasus baru stroke dari 70% penyakit stroke dan 87% kematian yang mendominasi pada masyarakat di wilayah atau negara *Global Burden of Disease* (GBD) dan indeks sosiodemografi (SDI) seperti Asia Tenggara, Asia Timur, dan Oseania dengan kondisi sosio-ekonomi wilayah yang rendah [8]. Sejumlah penelitian telah mengangkat kasus pengembangan model prediksi stroke untuk mempercepat identifikasi risiko stroke. Model tradisional seperti *Framingham Stroke Risk Profile* dan regresi logistik dirancang sebagai praktik klinis dengan penggabungan variabel-variabel sederhana dengan performa antarpopulasi serta keterbatasan dalam menangkap interaksi *non-linear* dan *high-dimensional features* [9].

Penelitian ini mengangkat data dari Kaggle yang berisi rekaman pasien dengan faktor klinis berjumlah 17 variabel X dan variabel Y sebagai label target. Dengan mengangkat dataset ini, pengujian pada penelitian ini akan memprediksi risiko stroke dengan menggunakan metode SGD dan mengevaluasi kontribusi kombinasi faktor-faktor klinis terhadap probabilitas seseorang yang mengalami stroke. *Stochastic Gradient Descent* (SGD) merupakan merupakan metode optimasi sederhana namun efisien untuk mencari nilai koefisien untuk meminimalkan *loss function* pada skala besar contohnya seperti data teks [10]. Dalam konteks penelitian ini, ada terdapat *baseline* model dalam prediksi basis *machine learning* yang akan sebagai tolok ukur yang menetapkan kinerja minimum yang harus dilampaui agar model memiliki kontribusi prediktif, karena tanpa baseline suatu model lain tidak bisa memiliki evaluatif yang jelas [11]. Perbandingan antara baseline dan SGD akan menunjang keunggulan SGD melalui mekanisme yang memungkinkan terjadinya pembaruan berdasarkan *gladien loss function*. Pertimbangan pada kedua model ini akan memberikan hubungan kompleks dan korelasi antar variabel klinis seperti risiko stroke.

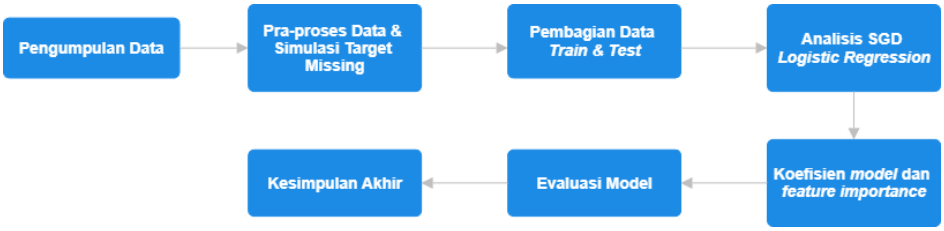
Selain itu, dapat dilihat bahwa berbagai metode prediksi telah dikembangkan, masih terdapat kesenjangan dalam hal efisiensi komputasi dan skalabilitas model pada dataset klinis berukuran besar. Metode tradisional berbasis regresi logistik konvensional memiliki akurasi prediksi yang moderat dan mengandalkan asosiasi statistik sederhana yang tidak mampu menangkap kompleksitas multifaktorial risiko stroke [9]. Di sisi lain, penggunaan algoritma berbasis *gradien* seperti SGD masih jarang diterapkan dalam konteks faktor klinis dan masih relatif jarang dibandingkan secara sistematis dengan *baseline model*, meskipun metode ini memiliki keunggulan dalam hal kecepatan konvergensi dan kemampuan menangani data skala besar. Di sisi lain, *baseline* juga masih terbatas dalam memprediksi variabel bersifat *non-linear* sehingga biasanya *baseline* hanya mencerminkan estimasi risiko minimum pada konteks data klinis yang heterogen.

Secara umum, penelitian ini bertujuan untuk membangun model prediksi risiko stroke berbasis SGD untuk mengidentifikasi variabel paling berpengaruh, *baseline* yang berperan sebagai acuan evaluatif akan menguji iteratif SGD, dan memprediksi risiko stroke berdasarkan penggabungan fitur klinis. Meskipun SGD umumnya digunakan pada dataset besar, penelitian ini mengevaluasi stabilitas, efisiensi iteratif, dan kemampuan seleksi fitur SGD terhadap peningkatan probabilitas stroke melalui mekanisme pembaruan gradien yang adaptif. Pendekatan ini menghadirkan perspektif baru mengenai kelayakan SGD sebagai optimisasi prediktif medis berbiaya rendah, relevan untuk populasi berpendapatan rendah yang memiliki beban stroke tinggi.

2. METODOLOGI PENELITIAN

Secara umum, penelitian ini berfokus pada pengembangan model prediksi risiko stroke berbasis 17 faktor klinis menggunakan metode SGD Regresi Logistik. Metode utama yaitu untuk mengoptimalkan fungsi *log-loss* dengan regresi logistik untuk klasifikasi biner risiko stroke, sehingga mampu menerapkan hubungan antara variabel klinis dan probabilitas stroke. Penerapan ini, bertujuan untuk memperoleh model prediksi yang mampu menangani kasus data klinis secara optimal.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786



Gambar 1. Alur Pengerjaan Penelitian

Gambar 1 menunjukan alur pengerjaan penelitian yang dimulai dari proses analisis tahap pra-pemrosesan data yang mencakup label *encoding* dan standardisasi fitur menggunakan standardScaler, dilanjut pembagian data training-testing dengan rasio 80:20 dan optimasi hyperparameter. Model yang telah dilatih akan dievaluasi menggunakan matriks klasifikasi medis standar untuk mengetahui performa prediksi yang akan membedakan kelompok berisiko dan tidak berisiko.

2.1 Sumber Data dan Variabel

Stroke adalah kelainan pada sistem serebrovaskular (pembuluh darah otak), yang ditandai dengan berkurang atau terhambatnya aliran darah dan oksigen ke otak, sehingga mengakibatkan kerusakan atau kematian jaringan otak dan gangguan fungsi otak [12]. WHO mencatat bahwa stroke menjadi penyebab kematian terbesar kedua di dunia [13]. Dataset ini berisi 35.000 data berdasarkan 17 faktor klinis yang berperan sebagai variabel prediktor (independen) dan 1 variabel target biner (dependen), dimana seluruh variabel tidak memiliki *missing value* karena tidak terdapat nilai yang hilang sehingga data bisa langsung diolah [14].

Variabel yang digunakan dalam penelitian ini terdiri dari 17 variabel independen yaitu Age (A), Gender (G), chest_pain (CA), high_blood_pressure (HBP), irregular_heartbeat (IH), shortness_of_breath (SOB), fatigue_weakness (FW), dizziness (DS), swelling_edema (SE), neck_jaw_pain (NJP), excessive_sweating (ES), persistent_cough (PC), nausea_vomiting (NV), chest_discomfort (CD), cold_hands_feet (CHF), snoring_sleep_apnea (SSA), dan anxiety_doom (AD), sedangkan dependen berupa stroke_risk_percentage (SRP), dan at_risk (AR). Adapun variabel penelitian disajikan dalam Tabel 1 sebagai berikut.

Tabel 1. Variabel Penelitian

Dependen		**Independen**															
SRP	**AR**	A	G	CA	HPB	IH	SOB	FW	DS	SE	NJP	ES	PC	NV	CD	CHF	SSA	AD
33.3	0	22	male	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0
100	1	52	male	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0
100	1	63	female	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0
44.5	0	41	male	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
84.8	1	53	male	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
18.3	0	22	male	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

2.2 Alur Penelitian

Bagian ini menguraikan tahapan metodologis yang diterapkan dalam menjawab permasalahan penelitian. Metode yang digunakan dirancang secara terstruktur agar proses analisis dapat dilakukan secara sistematis, objektif, dan selaras dengan tujuan penelitian. Mengembangkan alur pengerjaan penelitian untuk menghasilkan langkah analisis yang tepat.

2.2.1 Preprocessing Data

Dataset yang digunakan bersumber dari Kaggle dengan judul Stroke *Risk Prediction Dataset Based on Literature* yang dikompilasi pada tahun 2025 berdasarkan faktor risiko klinis dari berbagai literatur medis [14]. Data ini memuat 35.000 catatan pasien yang akan mengindikasikan keberadaan risiko stroke. Tahap awal dilakukan pemeriksaan kelengkapan data yang menunjukkan tidak terdapat *missing values* pada seluruh variabel sehingga dapat dilakukan transformasi data kategorik menggunakan label *Encoding* untuk mengubah variabel nominal seperti G menjadi numerik. Menjalani

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

penskalaan fitur menggunakan *StandardScaler* untuk menjamin semua fitur berada pada skala yang konsisten. Standardisasi ini penting untuk mempercepat konvergensi algoritma SGD. Mengingat ketidakseimbangan kelas yang signifikan pada kelas target (stroke), kami menerapkan Bobot kelas dalam model, dengan tujuan utama untuk meningkatkan sensitivitas terhadap kelas minoritas dan memaksimalkan *Recall*. Akhirnya, data yang diproses dibagi menjadi Data Pelatihan dan Data Pengujian dengan rasio 80:20 untuk tujuan validasi.

2.2.2 Data *Train* dan Data *Test*

Dataset yang telah melalui tahap pra-pemrosesan dibagi menjadi dua subset, yaitu *training set* dan *testing set* dengan proporsi 80:20. Pembagian ini bertujuan untuk memisahkan data yang digunakan dalam proses pelatihan model dengan data yang digunakan untuk evaluasi performa model secara independen. Data *train* berjumlah 28.000 observasi yang digunakan untuk melatih model SGD Regresi Logistik dan 7.000 observasi data *test* untuk menguji kemampuan model dalam memprediksi risiko stroke pada data. Dengan menggunakan *Stratified Split*, distribusi kelas data *train* dan *test* sekitar 63% untuk kelas tidak berisiko dan 37% untuk kelas berisiko, sehingga menghindari bias dalam evaluasi model.

2.2.3 SGD

SGD merupakan metode optimasi yang digunakan dalam pembelajaran mesin untuk meminimalkan fungsi kehilangan (*lost function*) melalui proses pembaruan parameter secara bertahap berdasarkan sampel data secara acak [15]. SGD dikenal karena kemampuannya menangani dataset besar dengan efisien [16]. Persamaan SGD dinyatakan sebagai berikut:

$$0_{t+1} = 0_{t-\eta}\nabla_0 L(0_t; x_1; y_1) \quad (1)$$

Keterangan :
0_t = Parameter model pada iterasi ke-t
η = Parameter model setelah pembaruan
0_{t+1} = Learning rate, mengatur besar langkah pembaruan
$L(0_t; x_1; y_1)$ = Gradien fungsi loss pada iterasi ke-t

Dalam implementasinya, penelitian ini menggunakan SGDClassifier dari library scikit-learn dengan konfigurasi fungsi "log_loss" secara otomatis mengaktifkan regresi logistik. Fungsi *log-loss* ini mengukur kesalahan prediksi probabilitas pada label aktual dan dioptimalkan melalui iterasi SGD. Optimasi hyperparameter menggunakan GridSearchCV dengan teknik 4-*fold cross-validation* untuk menemukan kombinasi parameter utama meliputi:
a. Alpha (= parameter regularisasi untuk mencegah *overfitting* dengan menambah *penalty* pada bobot model
b. *Learning rate* = strategi penyesuaian laju pembelajaran dengan metode optimal dan constant
c. Eta0 = nilai *learning rate* untuk langkah parameter
d. *Penalty* = jenis regularisasi dengan L1 (*Lasso Regression*), L2 (*Ridge Regression*), dan *elastic net*.

Dalam studi kasus ini digunakan satu sampel data dengan tiga fitur, yaitu $x_1 = 1{,}5$, $x_2 = 1{,}0$, dan $x_3 = 0{,}8$, dengan label aktual $y = 1$ yang menunjukkan kelas positif atau berisiko. Model diinisialisasi dengan parameter awal $\theta_0 = 0$, $\theta_1 = 0$, $\theta_2 = 0$, dan $\theta_3 = 0$, menggunakan *learning rate* $\eta = 0{,}01$ serta regularisasi $\alpha = 0{,}001$. Studi kasus ini digunakan untuk menunjukkan secara ringkas proses pembaruan parameter menggunakan metode *Stochastic Gradient Descent* (SGD) dalam memprediksi risiko berdasarkan data fitur yang diberikan untuk menghitung prediksi probabilitas, nilai loss, gradien, dan parameter baru setelah satu iterasi SGD..

1) Menghitung Prediksi

$$z = 0 + 0(1.5) + 0(1.0) + 0(0.8) = 0$$

$$\hat{y} = \frac{1}{(1 + e^0)} = 0.5$$

2) Menghitung Loss

$$L(\theta) = -[1 \times log(0.5)] = 0.693$$

3) Mengitung Gradien dan Update Parameter

$$\frac{\partial L}{\partial \theta_0} = (0.5 - 1) \times 1 + 0.001 \times 0 = -0.5$$

$$\theta_0' = 0 - 0.01 \times (-0.5) = 0.005$$

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

$$\frac{\partial L}{\partial \theta_1} = (0.5 - 1) \times 1.5 + 0.001 \times 0 = -0.75$$

$$\theta_1' = 0 - 0.01 \times (-0.75) = 0.0075$$

$$\frac{\partial L}{\partial \theta_2} = (0.5 - 1) \times 1.0 + 0.001 \times 0 = -0.5$$

$$\theta_2' = 0 - 0.01 \times (-0.5) = 0.005$$

$$\frac{\partial L}{\partial \theta_3} = (0.5 - 1) \times 0.8 + 0.001 \times 0 = -0.4$$

$$\theta_3' = 0 - 0.01 \times (-0.4) = 0.004$$

Hasil menunjukkan parameter θ₁ mengalami pertambahan terbesar (0.0075) karena nilai fitur $x_1 = 1.5$ paling tinggi, mengindikasikan fitur pertama memiliki kontribusi paling signifikan dalam pembelajaran model. Proses iterasi ini berlanjut untuk seluruh 28.000 sampel training dan diulang dalam beberapa epoch hingga konvergensi, menghasilkan parameter optimal untuk prediksi risiko stroke.

2.2.4 Koefisien Model dan Feature Importance

Precision-Recall Curve adalah alat yang digunakan untuk mengukur kinerja model klasifikasi, terutama pada kasus dimana kelas minoritas dalam dataset sangat tidak seimbang yang menggambarkan *trade-off* antara *precision* (ketepatan prediksi positif) dan *recall* (kemampuan mendeteksi semua kasus positif) pada berbagai *threshold* probabilitas [17]. Kurva ini sangat penting dalam konteks medis, khususnya untuk dataset dengan *class imbalance*, karena memberikan gambaran yang lebih informatif dibandingkan ROC *Curve* dalam mengevaluasi performa model pada kelas minoritas. *Precision* didefinisikan sebagai proporsi prediksi positif yang benar terhadap total prediksi positif:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Recall (sensitivitas) didefinisikan sebagai proporsi kasus positif yang berhasil dideteksi terhadap total kasus positif aktual:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Keterangan:
1) TP (*True Positive*) = jumlah kasus berisiko yang diprediksi benar
2) FP (*False Positive*) = jumlah kasus tidak berisiko yang salah diprediksi sebagai berisiko
3) FN (*False Negative*) = jumlah kasus berisiko yang tidak terdeteksi

Area Under Precision-Recall Curve (PR-AUC) digunakan sebagai metrik tunggal untuk mengevaluasi performa model secara keseluruhan. Nilai PR-AUC mendekati 1.0 menunjukkan bahwa model mampu mempertahankan *precision* tinggi sambil meningkatkan *recall*, yang berarti model dapat mendeteksi sebagian besar kasus berisiko dengan tingkat false alarm yang rendah.

2.2.5 Evaluasi Model

Metrik evaluasi yang digunakan pada penelitian klasifikasi medis meliputi *accuracy, precision, recall* (sensitivitas), F1-*score*, dan ROC–AUC. Metrik-metrik ini umum digunakan dalam penelitian kesehatan karena mampu menggambarkan keseimbangan antara prediksi positif dan negatif, serta akurasi model dalam mendeteksi kasus berisiko [18].

Model yang paling efektif kemudian dinilai secara independen melalui Data Uji untuk menentukan kemampuan prediksinya. Metrik kunci yang diterapkan untuk mengukur daya diskriminatif model adalah *Area Under the Curve* (AUC) dari ROC yang menunjukkan kemampuan yang cukup besar untuk membedakan antara klasifikasi stroke dan non-stroke di berbagai ambang batas . Selanjutnya, Metrik Kinerja Klasifikasi yang relevan, termasuk Presisi, *Recall*, F1-*Score*, dan Matriks Konfusi, digunakan untuk analisis menyeluruh. Perhatian khusus diarahkan pada Recall untuk kategori stroke, mengingat bahwa di bidang medis, mengurangi *False Negatives* (kasus stroke yang terlewatkan) merupakan tujuan utama.

3. HASIL DAN PEMBAHASAN

Pembahasan pada bab ini akan menyajikan hasil analisis data yang telah dilakukan serta pembahasan terhadap metode SGD. Penempatan model Regresi Logistik standar digunakan sebagai *baseline* untuk membandingkan kinerja model

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

berbasis SGD. Perbandingan ini bertujuan untuk menilai kontribusi penggunaan SGD terhadap efisiensi dan performa klasifikasi dalam memprediksi risiko stroke berdasarkan pemahaman yang jelas akan makna dan implikasi penelitian yang telah dilakukan.

3.1 SGD Regresi Logistik

Pada bagian ini dilakukan proses penentuan parameter terbaik dalam model SGD regresi logistik untuk memperoleh konfigurasi model yang paling optimal hubungan antara variabel independen dan dependen berdasarkan data Stroke Risk Prediction Dataset Based on Literature pada tahun 2025. Optimasi hyperparameter dilakukan pada GridSearchCV dengan 216 kombinasi parameter dan 4-*fold* validasi pada eksplorasi ruang. Parameter akan menentukan optimasi pada alpha (regularisasi), learning rate (strategi), eta0 (*learning rate* awal), dan *Penalty* (jenis regularisasi) yang disajikan pada tabel dibawah ini.

Tabel 2. Hyperparameter SGD Regresi Logistik

Hyperparameter SGD	
model_Alpha	0.001
model_eta0	0.001
model__l1_ratio	0.15

Tabel menunjukkan parameter yang digunakan pada model *Stochastic Gradient Descent* (SGD) regresi logistik dalam penelitian ini. Nilai alpha sebesar 0.001 digunakan sebagai parameter regularisasi untuk mengendalikan kompleksitas model dan membantu mengurangi potensi overfitting pada data berukuran besar. Nilai ini merepresentasikan tingkat penalti yang relatif ringan terhadap koefisien model. Parameter eta0 sebesar 0.001 digunakan sebagai nilai awal *learning rate* yang berfungsi mengatur besar langkah pembaruan bobot selama proses pelatihan. Nilai eta0 yang kecil memungkinkan proses pembelajaran berlangsung secara bertahap dan stabil, sehingga model dapat berkonvergensi dengan lebih terkendali. Nilai l1_ratio sebesar 0.15 menunjukkan proporsi kecil penggunaan regularisasi L1 dalam skema *elastic net*, dengan dominasi regularisasi L2. Kombinasi ini memungkinkan model tetap mempertahankan seluruh variabel prediktor tanpa melakukan eliminasi fitur secara agresif, sekaligus menjaga stabilitas koefisien dalam proses pembentukan model.

3.2 Koefisien Model dan Feature Importance

Hasil metode SGD regresi logistik memperlihatkan variabel independen yang mendominan pada masing-masing fitur klinis terhadap prediksi risiko stroke. Analisis ini dilakukan dengan mengevaluasi nilai koefisien absolut setiap variabel serta membedakan kelas positif dan negatif. Informasi yang disajikan pada Tabel 3 akan menyajikan 10 variabel tertinggi yang mendominan sebagai berikut.

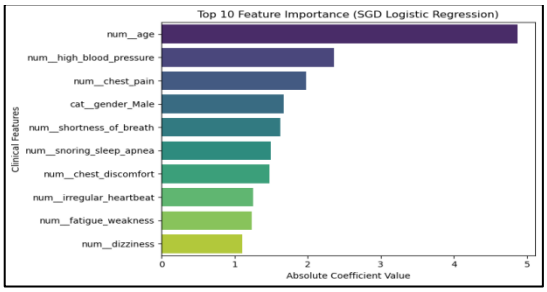
Tabel 3. Koefisien Regresi Logistik

	Feature	Coefficient	Abs_ Coefficient
0	num__age	4.868155	4.868155
2	num__high_blood_pressure	2.360180	2.360180
1	num__chest_pain	1.977221	1.977221
16	cat__gender_Male	1.670033	1.670033
4	num__shortness_of_breath	1.622389	1.622389
14	num__snoring_sleep_apnea	1.498227	1.498227
12	num__chest_discomfort	1.475878	1.475878
3	num__irregular_heartbeat	1.253288	1.253288
5	num__fatigue_weakness	1.233767	1.233767
6	num__dizziness	1.108530	1.108530

Age (usia) muncul sebagai prediktor terkuat dengan koefisien 4.868, hampir dua kali lipat dari fitur kedua yaitu High Blood Pressure (2.360). Hal ini konsisten dengan literatur medis yang menyatakan bahwa usia merupakan faktor risiko *non-modifiable* yang paling signifikan untuk penyakit kardiovaskular. Setiap pertambahan usia secara eksponensial meningkatkan probabilitas risiko. High Blood Pressure dan Chest Pain sebagai faktor kedua dan ketiga menegaskan bahwa kondisi kardiovaskular yang sudah terdiagnosis dan gejala klinis utama memiliki bobot prediksi yang sangat tinggi. Gender (Male) dengan koefisien 1.670 menunjukkan bahwa jenis kelamin laki-laki memiliki risiko lebih tinggi, sesuai dengan temuan epidemiologi bahwa pria memiliki onset penyakit kardiovaskular lebih dini dibanding wanita. Kelompok gejala seperti Shortness of Breath, Chest Discomfort, Irregular Heartbeat, dan Fatigue/Weakness semuanya memiliki koefisien positif yang substansial (di atas 1.1), mengindikasikan bahwa kombinasi gejala-gejala ini merupakan *warning signs* yang penting untuk deteksi dini.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

Snoring/Sleep Apnea dengan koefisien 1.498 menarik perhatian karena merupakan faktor yang sering diabaikan namun terbukti memiliki kontribusi signifikan terhadap risiko kardiovaskular, mendukung penelitian terkini tentang hubungan sleep disorders dengan kesehatan jantung.



Gambar 2. Visualiasi *Top Feature Importance*

Berdasarkan diagram batang yang disajikan pada Gambar 2, tampak bahwa avg_glucose_level (kadar glukosa rata-rata) dan age (usia) menduduki posisi teratas sebagai faktor penggerak yang paling signifikan. Ini mengindikasikan bahwa dalam model SGD Regresi Logistik yang telah dikembangkan, peningkatan pada kedua variabel tersebut berkaitan secara positif dengan peningkatan kemungkinan stroke. Dalam konteks klinis, hal ini sejalan dengan pengetahuan medis bahwa bertambahnya usia seseorang dan meningkatnya kadar gula dalam darahnya berpotensi meningkatkan risiko terjadinya gangguan pada pembuluh darah di otak.

Selanjutnya, variabel lain seperti bmi (indeks massa tubuh), hypertension (hipertensi), dan heart_disease (penyakit jantung) juga tampil sebagai elemen penting, meskipun dengan nilai koefisien yang lebih rendah dibanding usia dan glukosa. Diagram ini mengandalkan nilai absolut dari koefisien model untuk menetapkan urutan; dengan kata lain, fitur yang berada di urutan paling atas adalah fitur yang paling signifikan dalam memberikan informasi atau "sinyal" kepada model untuk membedakan antara individu yang memiliki risiko stroke dan yang tidak.

Dengan model regresi logistik yang digunakan, panjang batang pada diagram menunjukkan nilai log-odds. Fitur yang memiliki batang lebih panjang menunjukkan pengaruh matematis yang lebih signifikan dalam mengubah probabilitas prediksi ke arah kategori "berisiko". Dengan memahami visualisasi pada gambar 3.1, profesional medis atau analis dapat mengutamakan intervensi pada faktor-faktor penting seperti pengelolaan kadar gula dan pemantauan kesehatan pada populasi usia lanjut untuk pencegahan stroke yang lebih efektif.

3.3 Evaluasi SGD

Berdasarkan penentuan variabel yang paling berpengaruh, evaluasi antara kedua model menggunakan matrik akan menentukan nilai mana yang akan memprediksi secara komprehensif pada model *precision, recall*, f1-*score*, dan *accuracy*. Evaluasi dilakukan untuk membandingkan *Baseline* regresi logistik dan SGD regresi logistik pada data testing dengan jumlah 7.000 sampel dalam mengukur tingkat sensitivitas model dan prediksi pasien berisiko dan tidak berisiko stroke. Perbandingan model ini akan disajikan pada gambar dibawah ini untuk membedakan karakteristik kedua model tersebut.

```
=== SGD Logistic Regression ===
              precision    recall  f1-score

           0     0.9912    0.9704    0.9807
           1     0.9509    0.9853    0.9678

    accuracy                         0.9759
```

```
=== Baseline Logistic Regression ===
              precision    recall  f1-score

           0     0.9899    0.9742    0.9820
           1     0.9569    0.9829    0.9698

    accuracy                         0.9774
```

Gambar 3. *Classification Report* SGD Regresi Logistik dan *Baseline* Regresi Logistik

Gambar 3 menampilkan *classification report* lengkap dari kedua model yang menunjukkan metrik *precision*, *recall*, dan f1-*score* untuk setiap kelas. SGD Regresi Logistik (kiri) mencapai *accuracy* 97,59% dengan precision kelas 0 sebesar 99,12% dan recall kelas 1 sebesar 98,53%. *Baseline* Regresi Logistik (kanan) mencapai *accuracy* 97,74% dengan *precision* kelas 0 sebesar 98,99% dan *recall* kelas 1 sebesar 98,29%. Kedua model menunjukkan performa yang *excellent* dengan perbedaan yang minimal pada setiap matrik evaluasi.

Tabel 4. Perbandingan Hasil Klasifikasi

Model	Kelas	Precision	Recall	F1-Score	Accuracy

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

Baseline LR	0	0.9899	0.9742	0.9820	0.9774
	1	0.9569	0.9829	0.9698	
SGD LR	0	0.9912	0.9704	0.9807	0.9759
	1	0.9509	0.9853	0.9678	

Berdasarkan hasil evaluasi, model *Baseline* Regresi Logistik dan SGD Regresi Logistik menunjukkan performa klasifikasi yang sangat baik dan konsisten pada data pengujian. Model *Baseline* mencapai nilai akurasi sebesar 97,74%, sedangkan model SGD memperoleh akurasi 97,59%, dengan selisih 0,15%. Pada skenario pengujian dengan 7.000 sampel, perbedaan jumlah prediksi yang tidak tepat antara kedua model berada pada rentang yang sangat kecil, sehingga tidak merepresentasikan perbedaan performa yang bermakna secara praktis.

Pada kelas 0, model SGD mencatat nilai *precision* 0,9912, sedikit lebih tinggi dibandingkan *Baseline* 0,9899. Hal ini mengindikasikan bahwa proporsi prediksi negatif yang benar terhadap seluruh prediksi kelas 0 tetap sangat tinggi pada kedua model. Dari sisi recall, Baseline memperoleh nilai 0,9742, sedangkan SGD mencapai 0,9704, dengan selisih 0,38%. Perbedaan ini menunjukkan bahwa kemampuan kedua model dalam mengenali sampel yang benar benar termasuk kelas 0 tetap stabil dan hampir identik.

Pada kelas 1, *Baseline* Regresi Logistik memiliki nilai *precision* 0,9569, sedangkan SGD Regresi Logistik memperoleh *precision* 0,9509, selisih 0,60%. Meski *Baseline* menunjukkan precision sedikit lebih tinggi, kedua model tetap mempertahankan ketepatan prediksi positif yang kuat. Pada metrik recall, model SGD mencapai nilai 0,9853, sedikit lebih tinggi dibanding *Baseline* 0,9829, dengan selisih 0,24%. Nilai recall yang sangat tinggi pada kedua model menegaskan bahwa sensitivitas deteksi terhadap sampel berisiko tetap terjaga, termasuk pada model SGD.

Nilai F1-score pada masing masing kelas kembali menunjukkan hasil yang sangat berdekatan. Pada kelas 0, *Baseline* memperoleh F1 0,9820, sedangkan SGD 0,9807, dengan selisih 0,13%. Pada kelas 1, *Baseline* mencapai F1 0,9698, sementara SGD 0,9678, dengan selisih 0,20%. Kedekatan nilai ini menandakan bahwa model SGD tetap mampu mempertahankan keseimbangan *trade off* antara *precision* dan *recall* secara konsisten. Secara keseluruhan, kedua model memiliki kualitas prediksi yang setara, tanpa adanya indikasi penurunan performa yang substansial pada model SGD. Semua metrik berada dalam rentang yang sangat baik, sehingga kedua pendekatan dapat dinilai valid dan reliabel untuk klasifikasi risiko berbasis faktor demografis dan klinis pada data testing.

Tabel 5. *Model Evaluation Scores*

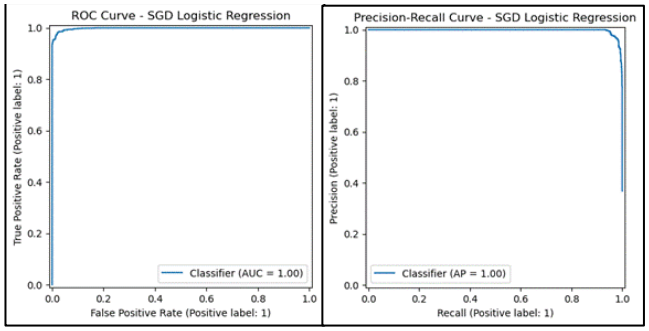
Confusion Matrix	
ROC-AUC	0.998405
PR-AUC	0.997485

Sesuai dengan Tabel 5, model ini menunjukkan kemampuan luar biasa dalam membedakan antara kelompok yang sehat (Kelas 0) dan kelompok yang berisiko (Kelas 1). Keberhasilan dalam memprediksi 4.292 kasus Negatif Benar menunjukkan spesifisitas sebesar 97,04%, yang berarti model ini sangat andal dalam mengidentifikasi orang-orang yang tidak menderita kondisi medis, sehingga mengurangi kebutuhan untuk menjalani tes tambahan yang tidak diperlukan. Sebaliknya, kemampuan untuk mengidentifikasi 2.539 kasus Positif Benar dengan tingkat keberhasilan 98,53% menandakan bahwa model ini sangat efisien sebagai alat deteksi. Angka ini memberikan tingkat kepercayaan yang tinggi bagi tenaga medis bahwa sebagian besar pasien yang benar-benar sakit akan terdeteksi oleh sistem ini.

Meskipun jumlahnya relatif kecil, 38 kasus *False Negative* (1,47%) merupakan elemen yang paling penting untuk dianalisis dalam bidang medis. Kesalahan ini muncul ketika sistem memprediksi bahwa pasien berada dalam keadaan aman, padahal mereka sebenarnya menghadapi ancaman terhadap kesehatan. Dalam konteks klinis, diagnosis yang terlewat seperti ini dapat mengarah pada keterlambatan penanganan yang berpotensi mematikan. Namun, jika dilihat dari perspektif statistik, tingkat 1,47% menunjukkan angka kesalahan yang sangat minimal. Ini menunjukkan bahwa model tersebut memiliki *Sensitivity (Recall*) yang sangat tinggi (sekitar 98,53%), yang merupakan syarat krusial untuk alat skrining agar tidak mengabaikan pasien yang membutuhkan pertolongan.

Terdapat 131 insiden *False Positive* (2,96%), di mana model memberikan sinyal bahaya kepada individu yang sebenarnya sehat. Di ranah medis, kesalahan tipe ini sering kali disebut sebagai "*False Alarm*". Meskipun hal ini dapat menimbulkan kecemasan jangka pendek pada pasien dan meningkatkan beban administrasi untuk tes konfirmasi (seperti biopsi atau analisis laboratorium lebih lanjut), angka di bawah 3% dianggap sangat dapat diterima. Bagi perangkat skrining awal, lebih bijak untuk melakukan verifikasi ulang pada sejumlah kecil individu yang sehat daripada mengabaikan satu pasien yang memerlukan perhatian medis.

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786



Gambar 4. Perbandingan Model

Kedua model menghasilkan nilai ROC-AUC yang identik sebesar 0.9984, sangat mendekati nilai sempurna 1.0. Gambar 4 menunjukkan ROC *Curve* untuk SGD Regresi Logistik dengan kurva yang hampir vertikal di pojok kiri atas, mengindikasikan kemampuan diskriminasi yang *excellent* antara kelas positif dan negatif. ROC-AUC sebesar 0.9984 menunjukkan bahwa model memiliki probabilitas 99,84% untuk memberikan ranking lebih tinggi pada *sample* positif dibandingkan *sample* negatif yang dipilih secara acak. Ini mengkonfirmasi bahwa model memiliki separasi kelas yang sangat baik. PR-AUC sebesar 0.9975 mengkonfirmasi performa yang robust pada kelas minoritas (kelas 1 dengan proporsi 36,82%). Metrik PR-AUC lebih informatif dibandingkan ROC-AUC untuk kasus dengan *class imbalance*, dan nilai tinggi ini menunjukkan *trade-off precision-recall* yang optimal.

3.4 Diskusi Penelitian

Secara umum, penelitian ini berhasil membawa performa klasifikasi yang baik, tetapi dapat diketahui bahwa setiap metode memiliki beberapa keterbatasan yang perlu dipertimbangkan dalam menginterpretasikan hasil model. Pengakuan atas keterbatasan ini bukan menjadi titik kelemahan fundamental dari peneliti saja, melainkan kritis yang perlu dilakukan untuk mengetahui batas-batas validitas dan pengaplikasian terhadap konteks klinis. Tindakan ini akan bertujuan untuk membuka ruang diskusi dan memberikan arahan bagi penelitian selanjutnya akan hasilnya.

Pertama, kesamaan performa kedua model yang sangat tinggi mengindikasikan bahwa dataset mungkin terlalu *clean* atau problem terlalu mudah, *baseline regresi logistik* mencapai *accuracy* 97,74% dan SDG regresi logistik 97,59%, ini menimbulkan bahwa merefleksikan kompleksitas kasus real-world tidak selalu stabil dengan missing *values, outliers*, dan *noise*. Dataset stroke ini juga merupakan data sintetis yang dimungkinkan sudah mengalami pengolahan idealisasi yang bukan rekam medis elektronik sehingga sulit untuk melakukan variabilitas dan ketidakpastian data klinis. Pada bagian *missing values, outliers*, serta *noise* biasanya akan selalu terjadi pada data akibat keterbatasan pemeriksaan atau pasien yang tidak teratur, kemudian kesalahan pengukuran atau dokumentasi yang tidak konsisten akan pasti terjadi pada setiap dataset, tetapi data ini minim atau bersih pada kondisi seperti ini.

Kedua, model hanya menggunakan 17 fitur input sedangkan *clinical assessment* komprehensif melibatkan banyak parameter tambahan termasuk *lab values* (cholesterol, troponin, HbA1c), *imaging results*, dan *detailed medical history*. Keterbatasan fitur ini dapat mempengaruhi generalizability model untuk kasus-kasus kompleks yang memerlukan integrasi data multimodal dari berbagai sumber diagnostik. Tidak tersedianya data seperti riwayat perubahan tekanan darah atau gejala dalam menangkap dinamika risiko stroke yang dapat berubah seiring dengan perubahan gaya hidup pasien.

Ketiga, penelitian ini belum melakukan validasi eksternal pada dataset dari institusi atau populasi yang berbeda. Performance model mungkin berbeda pada *demographic* yang berbeda (usia, etnis, geographic location) atau healthcare setting dengan karakteristik pasien yang berbeda. Proses ini akan dikembangkan untuk membuktikan bahwa populasi Asia Tenggara yang memiliki tingkat stroke tinggi untuk mengingatkan perbedaan prevalensi faktor resiko.

Keempat, penelitian ini hanya membandingkan dua variasi regresi logistik. Perbandingan dengan algoritma machine learning lain seperti Random Forest, XGBoost, atau Neural Networks akan memberikan perspektif yang lebih lengkap tentang performance ceiling untuk problem ini. Dapat ditemukan pada kasus-kasus pasien yang memiliki resiko atipikal yang mungkin misclassified pada model linear.

4. KESIMPULAN

Penelitian ini menunjukkan bahwa penerapan model *Stochastic Gradient Descent*(SGD) untuk Regresi Logistik dan Regresi Logistik Baseline menghasilkan hasil yang sangat mengesankan dan hampir serupa dalam mengidentifikasi risiko stroke di tahap awal. Dengan tingkat akurasi yang berkisar antara 97,59% hingga 97,74% serta nilai ROC-AUC yang sangat tinggi mencapai 0,9984, kedua model memperlihatkan kemampuan diskriminasi yang sangat baik dalam menggolongkan individu yang berisiko. Penilaian terhadap signifikansi fitur menunjukkan bahwa variabel klinis seperti

JURNAL KOMPUTER TEKNOLOGI INFORMASI SISTEM KOMPUTER
p-ISSN: 2962-3022 | e-ISSN: 2963-7104
Volume 4 No 3 Februari 2026
Hal: 2028-2037
http://ejurnal.lkpkaryaprima.id/index.php/juktisi
DOI: https://doi.org/10.62712/juktisi.v4i3.786

usia, hipertensi, dan nyeri dada adalah faktor prediktor utama yang memiliki pengaruh paling signifikan terhadap model, dan hasil ini sangat mendukung temuan dalam literatur medis yang ada. Yang paling penting dalam konteks kesehatan adalah kemampuan model untuk mencapai tingkat *false negative*yang sangat rendah, yakni hanya sekitar 1,47%. Rendahnya tingkat kegagalan deteksi ini menegaskan bahwa model SGD memiliki sensitivitas yang sangat tinggi (98,53% VS 98,29%) dan *false negative*lebih rendah, menjadikannya alat yang andal untuk digunakan sebagai sarana skrining awal dalam mengurangi risiko kematian serta kecacatan permanen akibat keterlambatan dalam menangani pasien stroke pada pendekatan yang dikembangkan menjadi screening tool awal resiko stroke, tetap membutuhkan validasi sebagai keamaan praktik medis. Berdasarkan hasil yang diperoleh, penelitian selanjutnya disarankan untuk melaksanakan validasi eksternal dengan menggunakan dataset dari populasi yang lebih beragam, baik secara geografis maupun demografis, agar dapat memverifikasi tingkat generalisasi model yang lebih kokoh sebelum implementasi secara luas. Mengingat bahwa pengoptimalan hyperparameter secara mendalam pada dataset statis ukuran sedang tidak menunjukkan peningkatan kinerja yang substansial dibandingkan dengan model awal, penelitian mendatang dapat difokuskan pada memanfaatkan keunggulan khusus SGD dalam situasi pembelajaran online. Ini memungkinkan model untuk secara berkelanjutan belajar dalam waktu nyata dari aliran data pasien yang masuk ke rumah sakit tanpa perlu melatih kembali model dari awal. Selain itu, pengembangan antarmuka sistem dukungan keputusan yang terintegrasi langsung ke dalam alur kerja klinis di fasilitas kesehatan perlu dilakukan. Langkah ini sangat penting untuk memastikan bahwa hasil prediksi model dapat dipahami dengan baik oleh tenaga kesehatan dan memberikan kontribusi nyata dalam pengambilan keputusan preventif yang tepat waktu untuk pasien.

REFERENCES

[1] V. L. Feigin and others, “Global burden of stroke and risk factors in 2021 and beyond,” *Lancet Neurol.*, vol. 24, no. 2, pp. 123–134, 2025, doi: 10.1016/S1474-4422(24)00401-5.
[2] A. Forrester and others, “Stress hyperglycemia and outcomes in acute stroke,” *Stroke*, vol. 51, no. 6, pp. 1831–1838, 2020, doi: 10.1161/STROKEAHA.119.028343.
[3] A. Khosla and others, “Blood pressure variability and cerebrovascular damage,” *Hypertension*, vol. 76, no. 2, pp. 375–383, 2020, doi: 10.1161/HYPERTENSIONAHA.120.14821.
[4] Y. Chen and others, “Clinical predictors of ischemic stroke in young adults,” *Neurology*, vol. 96, no. 15, pp. e1956--e1966, 2021, doi: 10.1212/WNL.0000000000011834.
[5] E. Harshfield and others, “Obesity, vascular dysfunction, and stroke risk,” *Circ. Res.*, vol. 128, no. 4, pp. 512–524, 2021, doi: 10.1161/CIRCRESAHA.120.317987.
[6] C. S. Katsouras, X. M. Sakellariou, A. Bechlioulis, and L. Lakkas, “Current insights and challenges in the management of left ventricular thrombus,” no. xxxx, p. 2025, 2025.
[7] M. R. Zemaitis, J. M. Boll, and M. A. Dreyer, “Peripheral Arterial Disease,” *StatPearls [Internet]*, 2023, [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK430745/
[8] World Stroke Organization, *World Stroke Report 2022*. Geneva, Switzerland: World Stroke Organization, 2022. [Online]. Available: https://www.world-stroke.org
[9] J. Chahine and others, “Limitations of traditional stroke risk prediction models,” *J. Stroke Cerebrovasc. Dis.*, vol. 32, no. 4, 2023, doi: 10.1016/j.jstrokecerebrovasdis.2023.106996.
[10] R. Dwiyansaputra, G. S. Nugraha, F. Bimantoro, and A. Aranta, “Deteksi Sms Spam Berbahasa Indonesia Menggunakan Tf-Idf Dan Stochastic Gradient Descent Classifier,” *J. Teknol. Informasi, Komput. dan Apl.*, vol. 3, no. 2, pp. 200–207, 2021.
[11] J. Klein, S. Bhulai, and R. Van Der Mei, “arXiv : 2301 . 03318v1 [cs . LG] 9 Jan 2023 The Optimal Input-Independent Baseline for Binary Classification : The Dutch Draw,” pp. 1–16.
[12] Aulyra Familah, Arina Fathiyyah Arifin, Achmad Harun Muchsin, Mochammad Erwin Rachman, and Dahliah, “Karakteristik Penurunan Kebugaran Kardiorespirasi Pada Penderita Stroke Iskemik dan Stroke Hemoragik,” *Fakumi Med. J. J. Mhs. Kedokt.*, vol. 4, no. 6, pp. 456–463, 2024.
[13] World Health Organization, “Stroke: Key Facts,” Geneva, Switzerland, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/stroke
[14] M. A. Tusher, “Stroke Risk Prediction Dataset V2,” 2022, *Kaggle*.
[15] A. T. N. Hartono and H. D. Purnomo, “Pengembangan Stochastic Gradient Descent dengan Penambahan Variabel Tetap,” *J. JTIK (Jurnal Teknol. Inf. dan Komunikasi)*, vol. 7, no. 3, pp. 359–367, 2023, doi: 10.35870/jtik.v7i3.840.
[16] D. Dedy, “Klasifikasi Jenis Jamur Berdasarkan Citra Gambar Menggunakan Algoritma Stochastic Gradient Descent,” *Data Sci. Indones.*, vol. 4, no. 2, pp. 1–9, 2024, doi: 10.47709/dsi.v4i2.5014.
[17] J. -, S. Usman, and F. Aziz, “Analisis Perilaku Pelanggan menggunakan Metode Ensemble Logistic Regression,” *J. Teknol. Dan Ilmu Komput. Prima*, vol. 6, no. 2, pp. 90–97, 2023, doi: 10.34012/jutikomp.v6i2.4258.
[18] T. Saito and M. Rehmsmeier, “The precision-recall plot is more informative than the ROC plot,” *PLoS One*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0118432.