Vol. 5, No. 2, Agustus 2025, Hal. 59-67

Jurnal Software Engineering and Information System (SEIS)

https://ejurnal.umri.ac.id/index.php/SEIS/index

e-ISSN: 2809-0950

PEMODELAN *MACHINE LEARNING* DENGAN ALGORITMA *RANDOM FOREST* DALAM MEMPREDIKSI RISIKO *STROKE*

Doni Arman[1)], Nurul Sakhila Indayana[2*)], Finanta Okmayura[3)], Suci Putri Anjani[4)], Fitri Nur Dayani[5)], Muhammad Farhan[6)], Ariya Faturrahman[7)]

[1234567] Fakultas Matematika dan Ilmu Pengetahuan Alam (FMIPA), Universitas Riau

email: doniarmans6086@student.unri.ac.id[1], nurul.sakhila4321@student.unri.ac.id[2]*, finanta.okmayura@lecturer.unri.ac.id[3], suci.putri0132@student.unri.ac.id[4], fitri.nur0134@student.unri.ac.id[5], muhammad.farhan0123@student.unri.ac.id[6], ariya.faturrahman0116@student.unri.ac.id[7]

**Corresponding Author*

Abstract

Stroke is one of the diseases that significantly affects health and economy, becoming the second most common cause of death in the world after coronary heart disease. Based on data from the World Health Organization (WHO), stroke is ranked second as the leading cause of death in the world after ischemic heart disease. In 2019, stroke was responsible for around 11% of total global deaths. One important way to reduce the death rate from stroke is to make prevention efforts through early prediction. Machine learning methods, especially Random Forest, are used in this study to predict the risk of stroke. The data used comes from a public dataset that includes age, gender, blood pressure, blood sugar, smoking status, and other medical history. The research process includes data pre-processing stages (data cleaning, outlier handling, and category coding), model training using the Random Forest algorithm, and model evaluation using a confusion matrix to evaluate accuracy, precision, recall, and F1 score. The evaluation results show an accuracy value of 97.55%, which indicates very good predictive performance so that this model has very good predictive performance.

Keywords: *Machine Learning, Random Forest, Stroke*

Abstrak

Stroke adalah salah satu penyakit yang mempengaruhi kesehatan dan ekonomi secara signifikan, menjadi penyebab kematian kedua tersering di dunia setelah penyakit jantung koroner. Berdasarkan data dari *World Health Organization* (WHO), *stroke* menempati peringkat kedua sebagai penyebab utama kematian di dunia setelah penyakit jantung iskemik. Pada tahun 2019, *stroke* bertanggung jawab atas sekitar 11% dari total kematian global. Salah satu cara penting untuk mengurangi angka kematian akibat *stroke* adalah dengan melakukan upaya pencegahan melalui prediksi dini. Metode pengajaran mesin, terutama *Random Forest*, digunakan dalam penelitian ini untuk memprediksi risiko penyakit *stroke*. Data yang digunakan berasal dari dataset publik yang mencakup usia, jenis kelamin, tekanan darah, gula darah, status merokok, dan riwayat kesehatan lainnya. Proses penelitian mencakup tahap pra-pemrosesan data (pembersihan data, penanganan *outlier*, dan pengkodean kategori), pelatihan model menggunakan algoritma *Random Forest*, dan evaluasi model menggunakan *confusion matrix* untuk mengevaluasi akurasi, presisi, *recall*, dan skor F1. Hasil evaluasi menunjukkan nilai akurasi sebesar 97,55%, yang menandakan kinerja prediktif yang sangat baik sehingga model ini memiliki performa prediksi yang sangat baik.

Kata Kunci: *Machine Learning, Random Forest, Stroke*

PENDAHULUAN

Stroke merupakan masalah kesehatan serius, terutama pada lansia, karena dapat menyebabkan kerusakan otak permanen akibat berkurangnya suplai oksigen dan nutrisi. Berdasarkan data dari *World Health Organization* (WHO), *stroke* menempati peringkat kedua sebagai penyebab utama kematian di dunia setelah penyakit jantung iskemik. Pada tahun 2019, *stroke* bertanggung jawab atas sekitar 11% dari total kematian global (WHO, 2020). Kondisi ini menjadikan *stroke* sebagai penyakit mematikan yang membutuhkan perhatian serius dalam hal pencegahan dan penanganannya.

Di Indonesia, *stroke* menjadi risiko tinggi yang sering dipicu oleh penyakit penyerta seperti diabetes dan hipertensi. Dampak dari *stroke* sangat besar, tidak hanya menyebabkan kecacatan jangka panjang, tetapi juga meningkatkan beban ekonomi dan sosial masyarakat. Jika tidak segera dicegah dan ditangani, *stroke* dapat menyebabkan kematian mendadak atau gangguan fungsi tubuh yang menetap.

Maka dari itu, upaya deteksi dini sangat penting untuk mengurangi angka kematian dan kecacatan akibat *stroke*. Salah satu pendekatan yang kini banyak digunakan dalam bidang kesehatan adalah penerapan teknologi kecerdasan buatan, khususnya algoritma machine learning, untuk memprediksi risiko *stroke*. Studi ini menggunakan algoritma *Random Forest* dalam menganalisis data medis dalam jumlah besar, meliputi variabel kesehatan seperti usia, jenis kelamin, pendapatan, tingkat pendidikan, dan jenis pekerjaan.

Random Forest dipilih karena kemampuannya dalam menangani banyak variabel sekaligus serta menghasilkan prediksi yang stabil dan akurat. Selain membangun model prediksi, penelitian ini juga bertujuan untuk mengidentifikasi faktor-faktor risiko utama yang berkontribusi terhadap kejadian *stroke*. Hasil dari penelitian ini diharapkan dapat dikembangkan menjadi sistem pendukung keputusan bagi tenaga medis dalam melakukan pencegahan dan diagnosis dini *stroke*.

Evaluasi performa model dilakukan menggunakan metrik klasifikasi seperti akurasi, presisi, *recall*, dan F1-*score* yang dianalisis melalui *confusion matrix*. Dengan pendekatan ini, diharapkan hasil penelitian dapat memberikan kontribusi nyata dalam meningkatkan upaya deteksi dini, diagnosis, dan pencegahan *stroke* di lingkungan klinis berbasis teknologi data.

TINJAUAN PUSTAKA

1. *Stroke*

Stroke adalah gangguan aliran darah ke otak yang menyebabkan kerusakan jaringan otak. Dapat bersifat akut (terjadi mendadak) atau *kronik* (berkembang lambat). Menurut WHO, *stroke* menjadi penyebab utama kematian, terutama di negara berpenghasilan tinggi. Paling sering terjadi pada usia 55–64 tahun dan terkait dengan penyakit seperti hipertensi dan TIA.

2. *Machine Learning* dalam Dunia Medis

Machine Learning (ML) memungkinkan komputer belajar dari data masa lalu untuk membuat prediksi. Dalam dunia medis, ML digunakan untuk menganalisis data pasien dan membantu diagnosis, termasuk prediksi *stroke*, dengan akurasi yang tinggi meskipun data medis sering kompleks dan beragam.

3. Algoritma *Random Forest*

Machine Learning (ML) memungkinkan komputer belajar dari data masa lalu untuk membuat prediksi. Dalam dunia medis, ML digunakan untuk menganalisis data pasien dan membantu diagnosis, termasuk prediksi *stroke*, dengan akurasi yang tinggi meskipun data medis sering kompleks dan beragam.

$$Gini\ (D) = 1 - \sum_{i=1}^{m} \quad Pi^2 \qquad (1)$$

Yang mana:
Pi : Nilai peluang dari sebuah nilai tuple D pada suatu kelas
m : Jumlah label kelas

4. Penelitian Sebelumnya

Penelitian mengenai prediksi risiko *stroke* dengan metode *machine learning* telah banyak dilakukan, namun masing-masing memiliki fokus dan keterbatasan yang berbeda. Patmawati (2023), misalnya, membandingkan berbagai algoritma seperti *Logistic Regression*, *Decision Tree*, *Random Forest*, KNN, *Naive Bayes*, dan SVM untuk memprediksi stroke. Dari hasil perbandingan, SVM terlihat unggul dengan akurasi dan ROC-AUC mencapai 100%. Akan tetapi, temuan ini dipandang kurang

representatif karena penelitian tersebut tidak mengatasi ketidakseimbangan data, yang justru merupakan masalah umum pada data medis. Ketidakseimbangan ini berpotensi membuat model terlalu bias terhadap kelas mayoritas sehingga hasilnya tidak sepenuhnya dapat diterapkan pada kondisi nyata.

Sementara itu, Putri (2024) menekankan efektivitas *Random Forest* dalam mendeteksi pola risiko *stroke*. Dengan menggunakan dataset publik dari UCI, penelitian ini mampu mencapai akurasi 99,4%, presisi sempurna (1,00), dan *recall* sebesar 0,99. Capaian ini menunjukkan bahwa *Random Forest* merupakan salah satu algoritma yang sangat andal. Namun demikian, penelitian ini tidak menjelaskan secara detail strategi pra-pemrosesan yang digunakan, khususnya terkait masalah data yang tidak seimbang, sehingga menimbulkan pertanyaan apakah performa yang tinggi tersebut bisa konsisten dalam skenario klinis yang lebih kompleks.

Dari berbagai penelitian tersebut, terlihat adanya beberapa celah penting yang perlu diperhatikan. Pertama, sebagian besar penelitian belum memberikan solusi nyata dalam bentuk antarmuka yang ramah pengguna, sehingga hasil prediksi sulit dimanfaatkan langsung oleh tenaga medis. Kedua, isu ketidakseimbangan data masih menjadi tantangan besar dan belum banyak dibahas secara mendalam, meskipun hal ini sangat memengaruhi keakuratan model. Ketiga, fokus evaluasi model umumnya masih terbatas pada metrik akurasi, padahal indikator lain seperti *precision*, *recall*, dan *F1-score* memiliki peran penting dalam konteks prediksi penyakit. Terakhir, kebanyakan penelitian menggunakan dataset publik internasional, sehingga masih jarang ada kajian yang mempertimbangkan karakteristik data medis lokal.

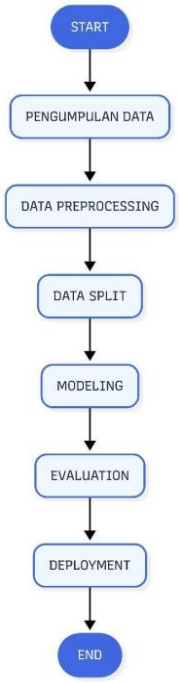
5. Kelebihan Menggunakan Metode *Random Forest*

Dalam penelitian ini, pemilihan metode yang tepat menjadi hal yang sangat penting untuk memperoleh hasil prediksi *stroke* yang akurat dan dapat diandalkan. Oleh karena itu, kami memilih algoritma *Random Forest* sebagai metode utama karena memiliki berbagai keunggulan yang sesuai dengan karakteristik data medis yang kompleks dan beragam. *Random Forest* dikenal memiliki akurasi tinggi, bahkan mencapai 95,2% dalam beberapa studi sebelumnya, serta mampu menangani dataset berukuran besar tanpa memerlukan proses seleksi variabel secara manual. Algoritma ini juga efektif dalam mengidentifikasi faktor-faktor risiko *stroke* secara akurat melalui proses klasifikasi yang kuat. Selain itu, *Random Forest* tahan terhadap *overfitting* karena menggunakan kombinasi banyak pohon keputusan dan sistem *voting* mayoritas. Kemampuannya dalam mengelola data yang heterogen dan kompleks menjadikannya pilihan yang tepat dalam konteks prediksi *stroke* berbasis data medis.

METODE PENELITIAN

1. Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional, membangun model prediksi *stroke* menggunakan algoritma *Random Forest*. Algoritma ini dipilih karena keunggulannya dalam menangani data kompleks dan kemampuannya mengurangi risiko *overfitting* melalui teknik *Ensemble Decision Tree*.



Gambar 1. Alur Penelitian

Penelitian ini mengikuti enam tahapan utama dalam membangun model prediksi *stroke* menggunakan algoritma *Random Forest,* yaitu:

a. Pengumpulan Data
   Data yang digunakan dalam penelitian ini diambil dari *platform Kaggle*, berisi

5110 data pasien dengan 12 atribut penting seperti usia, jenis kelamin, tekanan darah, kadar glukosa, riwayat hipertensi, dan lainnya. Dataset ini dipilih karena sesuai dengan kebutuhan analisis risiko *stroke* secara klinis dan komprehensif.

b. Data *Preprocessing*
Tahap ini mencakup pembersihan data, menghapus data duplikat, menangani nilai kosong (terutama pada atribut BMI yang memiliki 201 nilai kosong), serta memahami distribusi kelas. Ditemukan ketidakseimbangan data, yaitu 4.861 data non-*stroke* dan hanya 249 data *stroke*, sehingga perlu penanganan khusus seperti teknik *resampling* untuk menghasilkan model yang lebih adil.

c. Data *Split*
Data kemudian dibagi menjadi dua bagian, yaitu data latih (*training*) sebesar 80% dan data uji (*testing*) sebesar 20%. Data *split* adalah proses penting dalam pembelajaran mesin untuk memastikan bahwa model dapat dilatih pada sebagian data dan diuji pada data yang belum pernah dilihat sebelumnya, sehingga evaluasi performanya lebih objektif dan akurat.

d. *Modeling*
Pada tahap ini, algoritma *Random Forest* digunakan sebagai model utama. Model dilatih menggunakan parameter *default*, kemudian dilakukan *tuning hyperparameter* (*n_estimators, max_depth*) dengan *Random Search/Grid Search.*

e. Evaluation
Model dievaluasi menggunakan metrik evaluasi seperti akurasi dan AUC (*Area Under Curve*). Hasil menunjukkan performa tinggi, dengan nilai AUC mencapai 1.00, yang menunjukkan model sangat baik dalam membedakan antara kelas *stroke* dan non-*stroke*. Namun, hasil ini tetap dianalisis secara kritis dengan mempertimbangkan potensi *overfitting* dan ketidakseimbangan data.

f. Deployment
Model diimplementasikan menggunakan *Jupyter Notebook* dan *Google Colab*, sehingga dapat dijalankan secara interaktif dan fleksibel. Sistem juga dilengkapi antarmuka grafis (GUI) berbasis *Python* untuk *backend*, HTML untuk struktur tampilan, dan CSS untuk desain, yang dibangun menggunakan *Visual Studio Code.* Dengan ini, pengguna seperti tenaga medis dapat memasukkan data pasien dan langsung mendapatkan hasil prediksi risiko *stroke*.

2. Perangkat dan Alat yang Digunakan

Penelitian ini menggunakan bahasa pemrograman *Python* karena sintaksnya sederhana dan didukung banyak pustaka analisis data serta *machine learning*. *Platform* yang digunakan meliputi *Google Colab* untuk eksekusi kode berbasis *cloud* dan *Jupyter Notebook* untuk eksplorasi data secara interaktif. Dataset diperoleh dari *platform Kaggle*, berjumlah 5110 data pasien. Namun setelah dilakukan *preprocessing* seperti penghapusan nilai kosong, jumlah data yang digunakan dalam modeling menjadi 4869, dengan 12 atribut penting terkait risiko *stroke*. Untuk tampilan antarmuka, digunakan *Visual Studio Code* dengan kombinasi *Python* sebagai *backend*, HTML untuk struktur tampilan, dan CSS untuk desain, sehingga tenaga medis dapat memasukkan data pasien dan langsung memperoleh hasil prediksi melalui GUI yang interaktif dan mudah digunakan.

HASIL DAN PEMBAHASAN

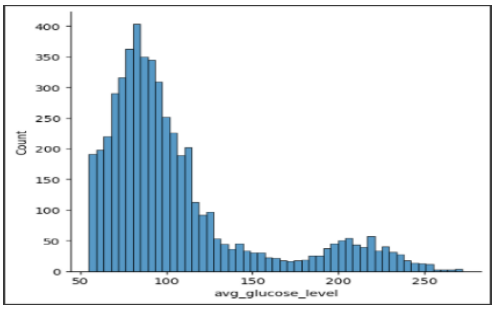
1. Persiapan Data

Tahap awal analisis data dimulai dengan memahami makna serta tujuan dari data yang digunakan, diikuti oleh evaluasi statistik deskriptif seperti nilai maksimum, rata-rata, median, dan standar deviasi untuk memberikan gambaran umum. Namun, keberadaan data yang hilang dapat memengaruhi keakuratan rata-rata, sehingga penting memastikan data bersih sebelum melanjutkan ke tahap berikutnya, yang umumnya tidak secara otomatis memeriksa nilai null. Visualisasi awal sering dilakukan dengan fungsi *head()* untuk melihat sebagian data dan *describe()* untuk menampilkan ringkasan statistik numerik. Dalam tahap pra-pemrosesan, data yang hilang sebaiknya diatasi melalui imputasi menggunakan nilai *mean, median,*

modus, atau nilai tertentu. Selain itu, atribut dengan banyak nilai unik (*high cardinality*) perlu dianalisis lebih lanjut karena dapat memengaruhi kinerja model, terutama jika terdapat pencilan (*outlier*) yang belum teridentifikasi, sehingga eksplorasi data lanjutan menjadi penting untuk meningkatkan kualitas dan akurasi model prediktif.

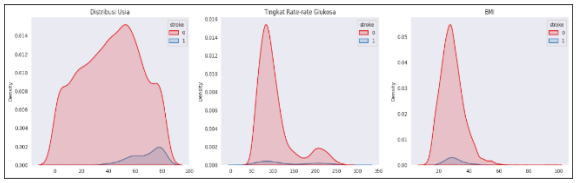
Attribute	Unique	Null
gender	3	0
age	104	0
hypertension	2	0
heart_disease	2	0
ever_married	2	0
work_type	5	0
Residence_type	2	0
avg_glucose_level	3979	0
bmi	418	201
smoking_status	4	0
stroke	2	0

Gambar 2. Jumlah nilai unik dan null di setiap kolom



Gambar 3. Visualisasi rata - rata glukosa

Dalam kenyataannya, atribut kadar glukosa rata-rata (*avg_glucose_level*), BMI, dan usia memiliki keterkaitan satu sama lain. Oleh karena itu, hubungan antara usia, AGL, dan BMI perlu dianalisis lebih lanjut. Hal ini juga menjelaskan mengapa atribut-atribut tersebut memiliki tingkat keunikan (*cardinality*) yang tinggi pada dataset.

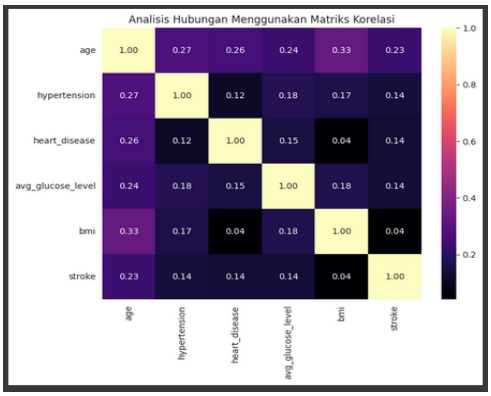


Gambar 4. Menyelidiki Usia, Rata-rata Glukosa dan BMI

Usia, BMI, dan kadar glukosa rata-rata merupakan atribut yang saling berkaitan dan menunjukkan peningkatan risiko terhadap diabetes maupun *stroke* seiring bertambahnya usia, sebagaimana terlihat pada grafik yang ditampilkan. Hal ini sekaligus menjelaskan tingginya nilai keunikan (*cardinality*) pada atribut-atribut tersebut. Korelasi yang cukup kuat antara risiko *stroke* dengan kadar glukosa dan BMI menjadikan ketiga atribut ini penting untuk diperhatikan dalam pengembangan model prediksi, baik dari aspek redundansi, keberadaan nilai kosong, kerusakan data, maupun proses pembersihannya. Kolom BMI, misalnya, mengandung sejumlah besar nilai kosong yang dapat memengaruhi keakuratan perhitungan statistik seperti rata-rata. Meskipun pengisian nilai kosong dengan rata-rata merupakan pendekatan umum, dalam kasus ini strategi yang lebih efektif adalah menghapus data kosong tersebut, mengingat jumlahnya mencapai 201 entri atau sekitar 10% dari total 4.908 data yang tersedia.

2. Data eksplorasi dan Analisis Data

Matriks korelasi merupakan cara yang sangat efisien untuk mengeksplorasi data baru. Metode ini digunakan untuk menentukan korelasi antara setiap pasangan kombinasi variabel dalam analisis hubungan.

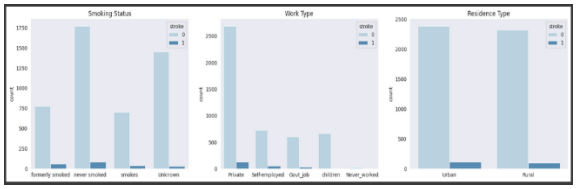


Gambar 5. Analisis Hubungan Menggunakan Matriks Korelasi

Nilai korelasi tersebut digunakan untuk mengukur seberapa kuat hubungan *linear* antara dua elemen dalam rekam medis elektronik pasien. Visualisasi korelasi ini ditampilkan dalam bentuk *colourmap* pada Gambar 5. Gambar 5 Analisis Hubungan Menggunakan Matriks Korelasi.

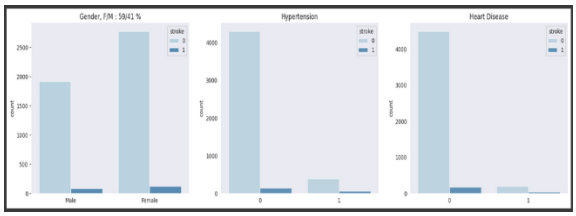
Nilai korelasi dimanfaatkan untuk menilai sejauh mana kekuatan hubungan linier antara dua atribut dalam data rekam medis elektronik pasien. Berdasarkan Gambar 5, mayoritas nilai korelasi berada di sekitar nol, yang menunjukkan bahwa hubungan antar variabel dalam dataset ini cenderung lemah. Dari temuan ini, beberapa asumsi awal terhadap data dapat mulai dibentuk. Misalnya, atribut seperti jenis pekerjaan (*work type*) dan jenis tempat tinggal (*residency type*)

diduga memiliki keterkaitan dengan kemungkinan seseorang mengalami *stroke*. Oleh karena itu, variabel-variabel tersebut akan dianalisis lebih mendalam pada tahap pemodelan selanjutnya.



Gambar 6. Bagan batang untuk pekerjaan, jenis tempat tinggal, dan kadar glukosa

Berdasarkan hasil visualisasi, tidak terlihat adanya hubungan yang relevan antara fitur-fitur seperti jenis pekerjaan, tempat tinggal, dan status merokok terhadap kejadian *stroke*. Oleh karena itu, fokus analisis diarahkan pada variabel lain seperti jenis kelamin, riwayat hipertensi, dan penyakit jantung yang secara klinis lebih berpotensi berkontribusi terhadap risiko *stroke*.



Gambar 7. Bagan batang untuk jenis kelamin, Hipertensi, penyakit jantung

Berdasarkan perbandingan tinggi antar batang pada grafik, tampak bahwa panjang masing-masing batang relatif serupa, yang mengindikasikan adanya suatu bentuk hubungan antar variabel. Selain itu, terlihat pula adanya ketimpangan distribusi pada fitur *stroke*, yang menunjukkan perbedaan proporsi yang tidak seimbang antara kategori yang ada.

3. Data *Prepocessing*

Pada tahap eksplorasi data, ditemukan bahwa data bersifat tidak seimbang (imbalanced). Dari keseluruhan data, hanya terdapat 209 pasien yang mengalami *stroke*, sedangkan 4699 pasien lainnya tidak mengalami *stroke*. Ketidakseimbangan ini menyebabkan model menghasilkan akurasi sebesar 96%, namun akurasi tersebut tergolong menyesatkan, karena model cenderung hanya mengklasifikasikan data ke kelas mayoritas, yaitu kelas 0 (tidak *stroke*). Oleh karena itu, dilakukan penanganan terhadap data yang tidak seimbang agar model dapat mengenali kedua kelas dengan lebih adil dan akurat.

```
Patient yang tidak punya stroke 0.04258353708231459
Patient yang punya stroke: 0.9574164629176855
stroke
0    4699
1     209
Name: count, dtype: int64
```

Gambar 8. Proporsi Pasien yang Mengalami dan Tidak Mengalami *Stroke*

Langkah berikutnya adalah melakukan encoding data kategorikal dengan *OneHotEncoder* dari sklearn, karena model machine learning tidak dapat memproses data dalam bentuk teks. Setelah itu, data dibagi menjadi data latih (*training set*) dan data uji (*testing set*), lalu dilakukan feature scaling untuk menyamakan skala antar fitur agar tidak ada yang mendominasi, sehingga model tetap stabil dan akurat.

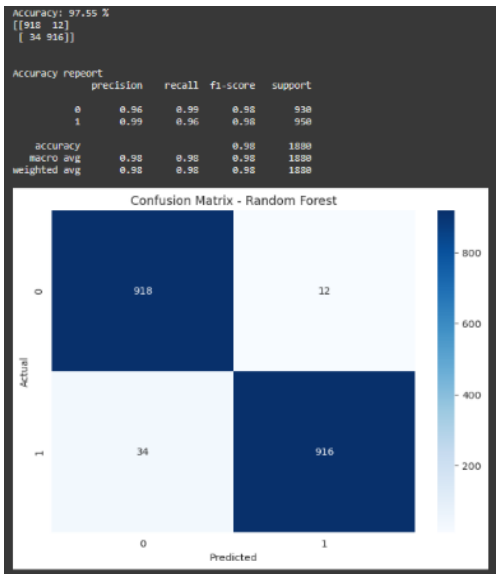
4. Model Pembelajaran Mesin Menggunakan *Random Forest*

Setelah melewati tahap pra-pemrosesan, pembangunan model *machine learning* dilakukan menggunakan algoritma *Random Forest.* Algoritma ini dipilih karena kemampuannya dalam menangani data berukuran besar, ketahanannya terhadap *overfitting*, serta kemampuannya menghasilkan prediksi yang konsisten dan stabil. *Random Forest* bekerja dengan membuat banyak pohon keputusan (*decision trees*) dari berbagai subset data, kemudian menggabungkan hasil prediksi tiap pohon melalui mekanisme *voting*, sehingga pendekatan *ensemble* ini dapat meningkatkan akurasi model.

Model dilatih menggunakan data latih yang telah diseimbangkan untuk menghindari bias, lalu diuji menggunakan data uji. Evaluasi performa dilakukan dengan sejumlah metrik seperti akurasi, presisi, *recall*, F1-*score*, dan AUC-ROC. Berdasarkan hasil evaluasi, model Random Forest menunjukkan kinerja yang sangat baik dalam memprediksi kemungkinan *stroke*, dengan tingkat akurasi yang mencapai lebih dari 95%.

5. Evaluasi Model dan Interpretasi Hasil

Model *Random Forest* yang telah dilatih dan diuji menunjukkan performa yang sangat baik dalam memprediksi kejadian *stroke*. Evaluasi dilakukan menggunakan beberapa metrik performa klasifikasi, yakni akurasi, presisi, *recall*, dan F1-*score*, yang dirangkum pada Gambar 9. Untuk memberikan gambaran yang lebih detail mengenai kinerja model dalam membedakan antara pasien yang mengalami stroke dan yang tidak, digunakan ***confusion matrix*** sebagaimana ditampilkan pada Gambar 9.

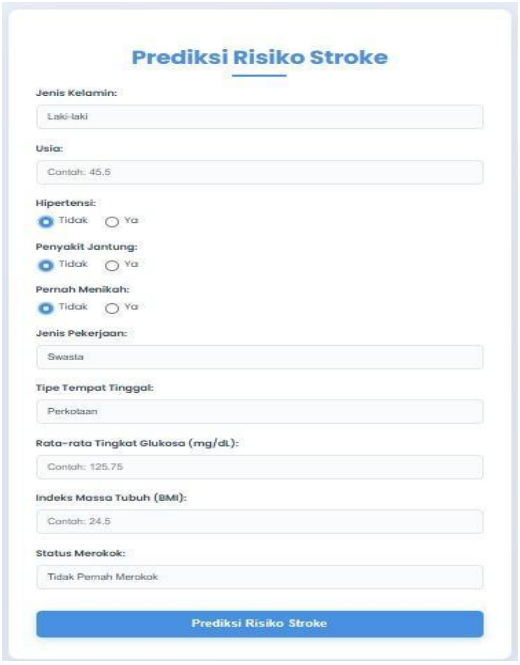


Gambar 9. *Confusion Matrix* Model *Random Forest*

Confusion matrix pada Gambar 4.4.1 menunjukkan bahwa model mampu mengenali sebagian besar pasien dengan tepat, baik pada kelas positif (*stroke*) maupun kelas negatif (tidak *stroke*). Hal ini ditunjukkan oleh tingginya jumlah *True Positive* (916) dan *True Negative* (918), sementara jumlah *False Positive* (12) dan *False Negative* (34) relatif rendah.

Model mencapai nilai akurasi sebesar 97,55%, yang menandakan kinerja prediktif yang sangat baik. Selain itu, nilai *standard deviation* (SD) dari validasi silang tercatat hanya sebesar 0,64, mengindikasikan konsistensi performa model pada data yang berbeda. Nilai AUC-ROC yang tinggi juga menegaskan kemampuan model dalam memisahkan antara dua kelas secara akurat.

Secara keseluruhan, hasil evaluasi ini menunjukkan bahwa algoritma *Random Forest* sangat potensial untuk diimplementasikan dalam sistem pendukung keputusan medis, khususnya untuk skrining dini terhadap risiko *stroke*.

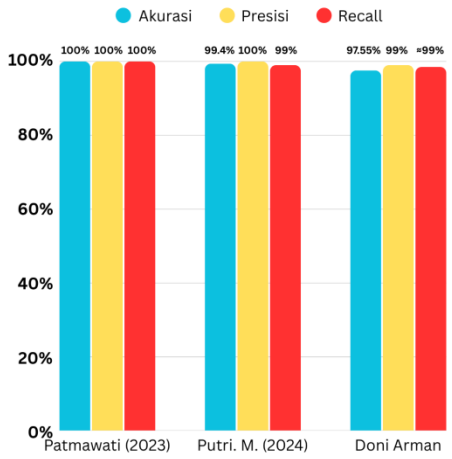


Gambar 10. Tampilan GUI

Gambar 11. Tampilan GUI

Alur aplikasi prediksi risiko *stroke* ditunjukkan pada Gambar 4.4.3 Tampilan GUI. Terdapat form untuk mengisi data kesehatan seperti usia, tekanan darah, BMI, dan riwayat medis. Setelah diisi dan dikirim, hasilnya menunjukkan prediksi risiko *stroke* dan saran kesehatan.

6. Perbandingan dengan Penelitian Sebelumnya



Gambar 12. Grafik perbandingan antar penelitian

Gambar di atas memperlihatkan perbandingan akurasi, presisi, dan *recall* dari tiga penelitian yang membahas prediksi risiko stroke menggunakan pendekatan *machine learning*. Penelitian oleh Patmawati (2023) menunjukkan hasil evaluasi model dengan nilai akurasi 100% presisi, dan *recall* masing-masing sebesar 1,00. Putri (2024) juga mencatatkan hasil yang sangat tinggi, yaitu akurasi 99,4%, presisi 1.00, dan *recall* 0,99. Sementara itu, penelitian ini menghasilkan akurasi sebesar 97,55%, dengan presisi dan *recall* masing-masing sebesar ≈0,99. Meskipun nilai-nilai tersebut sedikit lebih rendah dibandingkan dua penelitian sebelumnya, pendekatan yang digunakan dalam penelitian ini lebih menyeluruh, karena mencakup proses validasi silang, penanganan data tidak seimbang melalui *resampling*, serta evaluasi berbasis metrik AUC-ROC yang mencapai nilai sempurna (1,00). Selain itu, pengembangan antarmuka pengguna (GUI) menjadi keunggulan tambahan yang mendukung penerapan model dalam konteks dunia nyata. Dengan demikian, model yang dikembangkan dalam penelitian ini tidak hanya memiliki performa yang tinggi, tetapi juga unggul dari sisi kelayakan implementasi dan kesiapan penggunaan.

KESIMPULAN

Penelitian ini berhasil membangun model prediksi penyakit *stroke* dengan menggunakan algoritma *Random Forest* berbasis data publik dari *Kaggle*. Proses pengolahan data dilakukan secara menyeluruh, mulai dari eksplorasi, penanganan nilai hilang dan data tidak seimbang, hingga pelatihan dan evaluasi model. Hasil evaluasi menunjukkan bahwa model memiliki kinerja yang sangat baik, dengan akurasi sebesar 97,55%, presisi dan *recall* mendekati 0,99, serta nilai AUC sebesar 1,00, yang menandakan kemampuan tinggi dalam membedakan antara pasien yang berisiko *stroke* dan yang tidak. Atribut seperti usia, BMI, dan kadar glukosa terbukti berperan penting dalam klasifikasi risiko *stroke*. Algoritma *Random Forest* terbukti efektif dan stabil dalam menangani data medis yang kompleks, menjadikannya sangat potensial untuk diterapkan sebagai sistem pendukung keputusan dalam diagnosis dini *stroke*.

DAFTAR PUSTAKA

Ary Prandika Siregar, Dwi Priyadi Purba, Jojor Putri Pasaribu, & Khairul Reza Bakara. (2023). Implementasi Algoritma *Random Forest* Dalam Klasifikasi Diagnosis Penyakit Stroke. *Jurnal Penelitian Rumpun Ilmu Teknik*, *2*(4), 155–164. https://doi.org/10.55606/juprit.v2i4.3039

Ashfania, G. A. M., Prahasto, T., Widodo, A., & Warsokusumo, T. (2023). Penggunaan Algoritma *Random Forest* untuk Klasifikasi Berbasis Kinerja Efisiensi Energi pada Sistem Pembangkit Daya. *ROTASI*, 24(3), 14-21.

Fadli, M., & Saputra, R. A. (n.d.). *KLASIFIKASI DAN EVALUASI PERFORMA MODEL RANDOM FOREST UNTUK PREDIKSI STROKE Classification And Evaluation Of Performance Models Random Forest For Stroke Prediction*. *12*. http://jurnal.umt.ac.id/index.php/jt/index

Pandhita, G., Samino, & Bustami, M. (2017). Skor ICH-GS untuk prediksi prognosis pasien *stroke* perdarahan intraserebral di Rumah Sakit Islam Jakarta Pondok Kopi. *CDK*, *44*(12), 847–850.

Patmawati. (2023). *https://bufnets.tech https://doi.org/10.59688/bufnets BULLETIN OF NETWORK ENGINEER AND PREDIKSI PENYAKIT STROKE MENGGUNAKAN SUPPORT VECTOR MACHINE (SVM) STROKE PREDICTION*

USING A SUPPORT VECTOR MACHINE (SVM). https://doi.org/10.59688/bufnets

Putri, M. (2024). Prediksi Penyakit *Stroke* Menggunakan Machine Learning Dengan Algoritma Random Forest. *Jurnal Infomedia: Teknik Informatika*.

Ristyawan, A., Nugroho, A., & Amarya, T. K. (2025). *Optimasi Preprocessing Model Random Forest Untuk Prediksi Stroke*. *12*(1), 29–44.

Teknika, J., & Ria Supriyatna, A. (n.d.). Teknika 17 (1): 163-172 Prediksi Penyakit Diabetes Menggunakan Algoritma Random Forest. *IJCCS*, *x*, *No.x*, 1–5.

Wahyu Setiyo Aji, P. (2023). *Stroke Disease Prediction Using Random Forest Method [Prediksi Penyakit Stroke Menggunakan Metode Random Forest]*

Wahyu Setiyo Aji, P., & Dijaya, R. (2023). Prediksi penyakit *stroke* menggunakan metode Random Forest. *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, *4*(4), 916–924.

World Health Organization. (2020). *The top 10 causes of death*. Diakses pada 28 Juni 2025 dari: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death