



Klasifikasi Penyakit Stroke Menggunakan Random Forest

Kharisma Tri Octaviyani^{1*}, Kaila Aleisya², Letarman Lase³, Jonnes Febboys Zai⁴

^{1,2,3,4}Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang, Kota Tangerang Selatan, Indonesia

Email : ^{1*}khrrsmatr@gmail.com, ²kailaaleishya81@gmail.com, ³goodletarmanlase@gmail.com,
⁴jonnfebboyszai@gmail.com

(* : coressponding author)

Abstrak - Penelitian ini bertujuan untuk memprediksi risiko stroke menggunakan algoritma random forest dengan menggunakan data demografi dan perilaku. Ini memproses data seperti usia, jenis kelamin, kebiasaan merokok, dan status kesehatan untuk menciptakan model yang akurat guna mengidentifikasi potensi risiko stroke seseorang. Model ini menggunakan proses pelatihan dan pengujian pada kumpulan data besar untuk meningkatkan akurasi prediksi. Evaluasi model memberikan hasil yang sangat memuaskan dalam hal akurasi dan presisi dalam mengidentifikasi faktor risiko. Dengan hasil tersebut, penelitian ini memberikan kontribusi signifikan pada bidang prediksi medis. Model ini bertujuan untuk mendukung sistem peringatan dini bagi petugas kesehatan dan pasien berisiko.

Kata Kunci : Prediksi Stroke, Random Forest, Data Demografis, Faktor Risiko, Algoritma Machine Learning, Analisis Perilaku, Klasifikasi Kesehatan.

Abstract - This research aims to predict stroke risk using the Random Forest algorithm, utilizing demographic and behavioral data. It processes data such as age, gender, smoking habits, and health status to create an accurate model for identifying an individual's potential stroke risk. The model employs training and testing on a large dataset to enhance prediction accuracy. Model evaluation yields highly satisfactory results in terms of accuracy and precision in identifying risk factors. With these results, this research makes a significant contribution to the field of medical prediction. The model is intended to support early warning systems for healthcare professionals and at-risk patients.

Keywords : Stroke Prediction, Random Forest, Demographic Data, Risk Factors, Machine Learning Algorithm, Behavioral Analysis, Health Classification.

1. PENDAHULUAN

Stroke masih menjadi salah satu penyebab utama kematian dan kecacatan di seluruh dunia. Jutaan infeksi baru dan kematian tercatat setiap tahunnya. Deteksi dini dan prediksi stroke yang akurat dapat secara signifikan mengurangi angka kematian dan meningkatkan efektivitas intervensi terapeutik. Algoritma pembelajaran mesin, khususnya model prediktif yang menggunakan Random Forest, semakin mendapat perhatian dalam beberapa tahun terakhir karena kemampuannya menangani kumpulan data besar dan hubungan kompleks antar variabel. Algoritma ini terbukti efektif dalam mengklasifikasikan risiko stroke berdasarkan berbagai faktor seperti data demografi, kebiasaan perilaku, dan parameter klinis.

Algoritma Random Forest sangat cocok untuk prediksi medis karena menggabungkan beberapa pohon keputusan untuk menghasilkan model prediksi yang lebih kuat, yang meningkatkan akurasi dan keandalan. Penelitian seperti yang ada pada "Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm" telah menunjukkan bagaimana data demografis dan perilaku dapat digunakan secara efektif untuk memprediksi risiko stroke (MDPI, 2023). Selain itu, penelitian dalam "Analysis of Stroke Classification Using Random Forest Method" juga menyelidiki klasifikasi jenis stroke, dengan mengevaluasi berbagai parameter dan algoritma untuk menghasilkan diagnosis stroke yang lebih tepat (ILKOM, 2022).

Penelitian lain seperti "Stroke Disease Prediction Using Random Forest Method" dan "Random Forest-based Prediction of Stroke Outcome" juga mengkaji efektivitas algoritma ini dalam memprediksi hasil stroke, tidak hanya kemungkinan terjadinya stroke, tetapi juga hasil klinisnya, dengan mengevaluasi data klinis dan laboratorium (ResearchGate, 2023). Penelitian tentang klasifikasi stroke iskemik yang menggunakan ekstraksi fitur dari jaringan saraf konvolusional (CNN) semakin meningkatkan akurasi diagnosis stroke dengan menggabungkan data citra medis



dengan teknik pembelajaran mesin (Semantic Scholar, 2023).

Dengan meningkatnya kejadian stroke, penerapan algoritma Random Forest dan model pembelajaran mesin lainnya sangat penting untuk mengembangkan alat prediksi yang dapat membantu tenaga medis dalam membuat keputusan yang lebih baik. Penelitian ini bertujuan untuk memberikan pemahaman yang lebih dalam mengenai prediksi stroke, dengan fokus pada penggunaan algoritma Random Forest dalam menganalisis risiko dan hasil stroke.

2. METODE

Dalam penelitian ini, literatur yang relevan mengenai penggunaan algoritma Random Forest untuk prediksi dan klasifikasi stroke dikumpulkan melalui pencarian sistematis di berbagai platform ilmiah seperti MDPI, ResearchGate, Arxiv, dan Semantic Scholar. Proses pencarian dimulai dengan penggunaan kata kunci seperti "Prediction of Stroke Disease using Random Forest," "Stroke classification using Random Forest," dan "Stroke outcome prediction using Random Forest." Kami memilih hanya artikel yang dipublikasikan dalam jurnal ilmiah peer-reviewed untuk memastikan kualitas dan validitas informasi. Beberapa jurnal yang kami anggap relevan dengan topik ini adalah sebagai berikut:

- 1) Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm (2023) – MDPI
- 2) Analysis of Stroke Classification Using Random Forest Method (2022) - ILKOM Journal
- 3) Stroke Disease Prediction Using Random Forest Method [Prediksi Penyakit Stroke Menggunakan Metode Random Forest] (2023) – ResearchGate
- 4) Random Forest-Based Prediction of Stroke Outcome (2023) – Arxiv
- 5) Ischemic Stroke Classification Using Random Forests Based on Feature Extraction of Convolutional Neural Networks (2023) - Semantic Scholar

2.1. Pendekatan Penelitian

Penelitian ini mengkaji penggunaan algoritma random forest dalam prediksi dan klasifikasi stroke menggunakan pendekatan kualitatif dengan teknik tinjauan literatur sistematis. Pendekatan ini menggunakan teknik pembelajaran mesin, khususnya random forest, untuk menggabungkan data sekunder dari berbagai penelitian terkait topik prediksi stroke. Penelitian ini bertujuan untuk mengevaluasi efektivitas dan penerapan algoritma random forest dalam analisis faktor risiko dan klasifikasi stroke dengan meninjau berbagai penelitian yang telah dipublikasikan.

2.2. Sumber dan Kriteria Pemilihan Literatur

Literatur yang dipilih untuk penelitian ini berasal dari jurnal internasional yang terindeks di database terpercaya seperti MDPI, ResearchGate, Semantic Scholar, dan Arxiv. Kriteria pemilihan literatur mencakup penelitian yang menggunakan algoritma random forest dalam konteks prediksi dan klasifikasi stroke menggunakan data demografi, perilaku, dan klinis. Studi yang dipilih juga harus memberikan informasi tentang hasil evaluasi model prediktif dan penggunaan teknik lain seperti ekstraksi fitur dari citra medis dan data medis lainnya. Penelitian menggunakan metrik yang jelas seperti presisi, presisi, dan perolehan juga menjadi prioritas.

2.3. Strategi Pencarian Literatur

Pencarian literatur dilakukan melalui berbagai platform dan database online seperti MDPI, ResearchGate, Semantic Scholar, dan Arxiv. Penelusuran menggunakan kata kunci seperti "Random Forest stroke classification", "Machine learning stroke prediction", dan "Stroke prediction using Random Forest.". Artikel yang relevan kemudian dipilih berdasarkan kesesuaian dengan topik, kualitas dan keandalan sumber. Beberapa makalah dipilih berdasarkan referensi silang dari penelitian yang diterbitkan sebelumnya.



2.4. Proses Analisis Literatur

Proses analisis literatur dilakukan dengan membaca, mengklasifikasikan, dan membandingkan hasil dari berbagai sumber yang relevan. Setiap studi terpilih dianalisis untuk menilai bagaimana random forest digunakan dalam konteks prediksi dan klasifikasi stroke. Berbagai penelitian dibandingkan berdasarkan keakuratan model, metode evaluasi, dan keterbatasan penelitian. Selain itu, pendekatan yang digunakan untuk ekstraksi data baik dari citra medis maupun data klinis juga dianalisis untuk menentukan dampaknya terhadap hasil prediksi.

2.5. Pelaporan Hasil dan Sintesis

Hasil tinjauan literatur ini disajikan dalam bentuk sintesis naratif yang menggabungkan hasil berbagai penelitian tentang pemanfaatan random forest dalam prediksi stroke. Fokus utama dari cakupan ini adalah bagaimana random forest diterapkan pada data demografi, perilaku, dan data medis lainnya untuk memprediksi risiko stroke dan mengklasifikasikan jenis stroke. Hasil penelitian dikategorikan berdasarkan akurasi model serta manfaat dan tantangan dari setiap penelitian. Selain itu, juga disertakan rekomendasi untuk penelitian lebih lanjut berdasarkan temuan yang ada.

2.6. Keterbatasan

Keterbatasan dalam penelitian ini termasuk keterbatasan pada sumber literatur yang tersedia, terutama yang berfokus pada Random Forest dan stroke. Beberapa artikel yang relevan mungkin memiliki batasan dalam hal dataset yang digunakan, seperti ukuran dataset yang kecil atau keterbatasan dalam keberagaman sampel. Selain itu, tidak semua studi yang diulas memiliki evaluasi menyeluruh terhadap performa model di dunia nyata, dan banyak yang lebih fokus pada eksperimen berbasis simulasi atau data yang sangat terstruktur. Penelitian ini juga terbatas pada literatur yang ada di platform tertentu, sehingga ada kemungkinan beberapa penelitian relevan tidak terjangkau.

3. PEMBAHASAN

No	Nama Peneliti dan Tahun	Metode yang dibahas	Tujuan penelitian	Hasil yang didapat
1	Muhammad Firdaus Banjar, Irawati, Fitriyani Umar, dan Lilis, Nur Hayati, tahun 2022	Penelitian ini menggunakan metode Random Forest, sebuah teknik klasifikasi yang menggabungkan banyak pohon keputusan untuk membuat prediksi. Metode ini mengaplikasikan teknik bagging dan menggunakan atribut tertentu untuk klasifikasi stroke. Selain itu, metode SMOTE diterapkan untuk menangani ketidakseimbangan data	Penelitian ini bertujuan untuk menganalisis variabel yang mempengaruhi klasifikasi penyakit stroke. Penelitian ini juga bertujuan untuk menguji performa klasifikasi stroke dengan mengukur akurasi, presisi, recall, dan f1score. Selain itu, penelitian ini ingin mengetahui efek penggunaan jumlah pohon yang berbeda pada metode Random Forest	Hasil penelitian menunjukkan bahwa penggunaan 100 pohon pada metode Random Forest memberikan performa terbaik dengan akurasi 86,82%. Atribut usia, kadar gula darah, dan indeks massa tubuh ditemukan sebagai variabel penting dalam prediksi stroke. Meskipun akurasinya cukup tinggi, nilai f1-score yang rendah menunjukkan adanya ketidakakuratan dalam prediksi data minoritas stroke

2	Priyo Wahyu Setiyo Aji dan Suprianto, tahun 2022	Penelitian ini menggunakan metode Random Forest, yang merupakan kombinasi pohon klasifikasi independen melalui proses pemungutan suara. Beberapa tahapan dilakukan, termasuk pra-pemrosesan data, pemrosesan, dan evaluasi hasil. Data yang digunakan berasal dari Kaggle dan diolah untuk menghasilkan prediksi stroke dengan akurasi tinggi	Penelitian ini bertujuan untuk membangun sistem prediksi penyakit stroke dengan memanfaatkan algoritma Random Forest. Sistem ini diharapkan membantu diagnosis dini stroke bagi kalangan medis. Prediksi yang akurat diharapkan dapat mempercepat deteksi data penanganan stroke pada pasien	Hasil penelitian menunjukkan tingkat akurasi model sebesar 99% setelah melalui proses Randomized Search Cross Validation untuk penyetelan parameter. Meskipun hasil akurasi tinggi, penelitian selanjutnya disarankan untuk membandingkan metode ini dengan algoritma lain. Hal ini bertujuan untuk memperluas pemahaman dan pemilihan model yang tepat untuk prediksi penyakit stroke
3	Carlos Fernandez-Lozano, Pablo Hervella, Virginia Mato-Abad, Manuel Rodríguez-Yáñez, Sonia Suárez-Garaboa, Iria López-Dequidt, Ana Estany-Gestal, Tomás Sobrino, Francisco Campos, José Castillo, Santiago Rodríguez-Yáñez, dan Ramón Iglesias-Rey, tahun 2021	Penelitian ini menggunakan metode Random Forest untuk memprediksi hasil jangka panjang pasien stroke. Algoritma ini mengidentifikasi variabel penting yang berhubungan dengan mortalitas dan morbiditas pasien tiga bulan setelah stroke. Data dari pasien stroke iskemik (IS) dan hemoragik non-traumatik (ICH) diolah untuk menghasilkan model prediksi dengan area bawah kurva (AUC) yang tinggi	Tujuan dari penelitian ini adalah untuk mengembangkan model prediksi menggunakan algoritma Random Forest dalam memproyeksikan risiko mortalitas dan morbiditas pada pasien stroke. Penelitian ini juga berusaha untuk mengidentifikasi variabel klinis dan biokimia yang paling memengaruhi prediksi tersebut. Selain itu, penelitian ini ingin membantu pengambilan keputusan dalam penanganan pasien stroke secara lebih tepat dan individual	Hasil penelitian menunjukkan bahwa variabel NIHSS pada 24 dan 48 jam serta suhu aksila adalah prediktor utama untuk evolusi pasien dalam tiga bulan. Model ini menunjukkan AUC lebih dari 0,90 untuk prediksi mortalitas, khususnya pada pasien IS dan IS+ ICH, sementara ICH memiliki AUC yang lebih rendah. Penelitian ini menyarankan penggunaan variabel tambahan, seperti data pencitraan dan genetik, untuk meningkatkan akurasi prediksi
4	Glori Stephani Saragih,	Penelitian ini menggunakan	Tujuan utama penelitian ini adalah	Metode yang diusulkan

	Zuherman Rustam, Dipo Aldila, Rahmat Hidayat, Reyhan E. Yunus, Jacub Pandelaki (2020).	Convolutional Neural Network (CNN) untuk ekstraksi fitur gambar CT scan otak pasien stroke iskemik. Setelah fitur diekstraksi oleh CNN, algoritma Random Forest digunakan untuk klasifikasi, menggantikan lapisan sepenuhnya tersambung pada CNN. Penelitian ini menguji berbagai kombinasi jumlah pohon dan kriteria (Gini atau Entropy) untuk optimasi akurasi klasifikasi.	untuk mendeteksi stroke iskemik pada CT scan secara otomatis dan akurat. Penelitian ini berusaha meningkatkan akurasi diagnosis dengan menggunakan gabungan CNN dan Random Forest. Dengan metode ini, diharapkan dokter dan tenaga medis dapat mengidentifikasi dan merespon kasus stroke secara cepat dan tepat.	menghasilkan akurasi Klasifikasi stroke iskemik sebesar 100% pada pengujian dengan dataset 10% menggunakan 100-350 pohon dengan kriteria Gini atau Entropy. Akurasi tertinggi dicapai saat persentase data pengujian kecil, karena model belajar lebih banyak dari data pelatihan. Hasil ini menunjukkan bahwa kombinasi CNN dan Random Forest efektif untuk deteksi stroke iskemik
5	Olamilekan Shobayo, Oluwafemi Zachariah, Modupe Olufunke Odusami, Bayode Ogunleye (2023).	Penelitian ini menggunakan algoritma Random Forest untuk memprediksi penyakit stroke berdasarkan data demografis dan perilaku. Metode ini dibandingkan dengan algoritma Logistic Regression dan Decision Tree untuk mengevaluasi performanya. Dataset digunakan dari Kaggle dan Diproses menggunakan teknik oversampling SMOTE untuk mengatasi ketidakseimbangan data.	Tujuan penelitian adalah untuk mengembangkan model prediksi penyakit stroke yang akurat dan mudah diinterpretasikan. Penelitian ini bertujuan untuk mengidentifikasi faktor risiko utama yang berkontribusi terhadap kejadian stroke. Hasilnya diharapkan dapat membantu tenaga medis dalam pencegahan dan deteksi dini stroke.	Algoritma Random Forest mencapai akurasi tertinggi sebesar 94.11%, lebih baik dibandingkan Logistic Regression dan Decision Tree. Usia dan indeks massa tubuh (BMI) ditemukan sebagai prediktor signifikan utama kejadian stroke. Hasil ini menunjukkan bahwa Random Forest adalah model prediksi yang efektif untuk kejadian stroke.



4. KESIMPULAN

Studi ini menunjukkan bahwa algoritma random forest mempunyai potensi besar dalam memprediksi risiko stroke berdasarkan data demografi dan perilaku. Model yang dikembangkan dalam penelitian ini terbukti efektif dalam mengidentifikasi individu berisiko tinggi terkena stroke dengan akurasi tinggi. Dengan menggunakan berbagai variabel seperti usia, jenis kelamin, kebiasaan merokok, dan status kesehatan, penelitian ini akan berkontribusi secara signifikan terhadap pengembangan sistem peringatan dini yang dapat digunakan oleh para profesional kesehatan. Hasil ini membuka peluang untuk penerapan lebih lanjut dari algoritma random forest dalam diagnosis dan pencegahan stroke. Sebagai langkah berikutnya, penelitian lebih lanjut dapat berfokus pada penerapan model ini pada data yang lebih besar dan beragam guna meningkatkan akurasi pada tingkat populasi.

DAFTAR PUSTAKA

- Adelina, V., Ratnawati, D. E., & Fauzi, M. A. (2018). Klasifikasi tingkat risiko penyakit stroke menggunakan metode GA-Fuzzy Tsukamoto. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(9), 3015–3021.
- Fadlilah, M. S., Wihandika, R. C., & Rahayudi, B. (2019). Klasifikasi penurunan fungsi kognitif pasien stroke menggunakan metode klasifikasi Random Forest. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 3(3), 3005–3013.
- As Sarofi, M. A., Irfamah, I., & Mukarromah, A. (2020). Identifikasi Genre Musik dengan Menggunakan Metode Random Forest. *Jurnal Sains dan Seni ITS*, 9(1), 79–86.
- Agustin, E., Eviyanti, A., & Azizah, N. L. (2023). Deteksi Penyakit Epilepsi Melalui Sinyal EEG Menggunakan Metode DWT dan Extreme Gradient Boosting. *Jurnal Media Informatika Budidarma*, 7(1), 117–127.
- Neuhaus, A. A., et al. (2017). Neuroprotection in stroke: The importance of collaboration and reproducibility. *Brain*, 140, 2079–2092.
- Stein, J. *Essentials of Physical Medicine and Rehabilitation Chapter 159 Stroke*. Philadelphia: Elsevier, 2020.
- Heniwati, Thabrany, H. Perbandingan Klaim Penyakit Katastropik Peserta Jaminan Kesehatan Nasional di Provinsi DKI Jakarta dan Nusa Tenggara Timur tahun 2014. *Jurnal Ekonomi Kesehatan Indonesia*, 2006.
- Roger, V.L., et al. "Heart disease and stroke statistics—2011 update: A report from the American Heart Association." *Circulation*, 2011.
- Fadli, M., & Saputra, R. A. (2023). Klasifikasi dan Evaluasi Performa Model Random Forest untuk Prediksi Stroke. *Jurnal Teknik*, 12(2), 72–80. Universitas Pamulang.
- Elbagoury, B.M., et al. "Hybrid Stacked CNN and Residual Feedback GMDH-LSTM Deep Learning Model for Stroke Prediction." *Sensors*, 2023.