



Analisis Performa Algoritma *Machine Learning* pada Prediksi Penyakit *Cerebrovascular Accidents*

Robi Aziz Zuama¹, Syaifur Rahmatullah², Yuri Yuliani¹

¹Fakultas Teknik dan Informatika, Sistem Informasi, Universitas Bina Sarana Informatika, Jakarta, Indonesia

²Fakultas Teknologi Informasi, Teknik Informatika, Universitas Nusa Mandiri, Jakarta Indonesia

Email: robi.rbz@bsi.ac.id, yuri.yyi@bsi.ac.id, syaifur.syl@nusamandiri.ac.id

Email Penulis Korespondensi: robi.rbz@bsi.ac.id

Abstrak—Cerebrovascular Accidents (stroke) merupakan penyakit yang mengancam dan penyebab kematian dan kecacatan dan disabilitas di dunia, di Indonesia kasus orang terkena stroke setiap tahun meningkat. Penyakit stroke dapat dicegah dengan menjalankan faktor gaya hidup sehat, makan makanan bergizi dan beraktifitas fisik. Tujuan penelitian ini adalah membuat sebuah model prediksi penyakit stroke yang efektif, sistem menggunakan parameter dari faktor gaya hidup, faktor yang dapat dikendalikan seperti faktor risiko medis dan faktor-faktor yang tidak dapat dikendalikan. Empat algoritma pengklasifikasi di usulkan yaitu multi layer perceptron, KNN, Decission Tree dan Random Forest. Hasil menunjukkan bahwa algoritma pengklasifikasi dapat bekerja efektif dengan hasilkan nilai akurasi mencapar sempurna 99,99% pada tingkat validasi 10K-Fold Validation

Kata Kunci: Prediction, Cerebrovascular Accidents, Stroke, Multi Layer Perceptron, KNN, Decission Tree, Random Forest

Abstract—Cerebrovascular Accidents (stroke) are a disease that threatens and causes death and disability and disability in the world, in Indonesia the number of people affected by stroke is increasing every year. Stroke can be prevented by adopting a healthy lifestyle, eating nutritious food, and doing physical activity. The purpose of this study is to create an effective stroke prediction model, the system uses parameters from lifestyle factors, controllable factors such as medical risk factors, and uncontrollable factors. Four classification algorithms are proposed, namely multi-layer perceptron, KNN, Decision Tree, and Random Forest. The results show that the classification algorithm can work effectively by producing a perfect score of 99.99% accuracy at the 10K-Fold Validation level of validation.

Keywords: Prediction; Cerebrovascular Accidents; Stroke, Multi-Layer Perceptron; KNN; Decision Tree; Random Forest

1. PENDAHULUAN

World Stroke Organization melaporkan secara global ada lebih 13,7 juta kasus stroke baru, dan satu dari empat orang di duni dengan umur diatas 25 tahun mengalami stroke [1]. kementerian kesehatan Indonesia melalui hasil Risesdas melaporkan bahwa terdapat peningkatan kasus stroke sekitar 3,9% dari tahun 2013 sampai tahun 2018, tepatnya sekitar 7% pada tahun 2013 dan pada tahun 2018 mengalami peningkatan menjadi 10,9% berdasarkan diagnosis pada penduduk dengan usia ≥ 15 tahun, secara nasional prevalensi stroke di Indonesia berdasarkan kelompok umur sering terjadi pada kelompok umur 55-64 tahun atau sekitar 33,3% dan proporsi penderita paling sedikit pada kelompok umur 15-24 tahun [2].

Cerebrovascular Accidents (stroke) menduduki peringkat kedua penyebab kematian setelah penyakit jantung iskemik, dan stroke juga merupakan penyebab utama kecacatan atau disabilitas, dan setiap tahun jutaan pasien stroke harus beradaptasi dengan kehidupan sehari-harinya untuk bertahan hidup untuk bergantung pada dukungan orang lain [3]. Penyebab utama stroke adalah penyumbatan atau pecahnya pembuluh darah di otak yang mengakibatkan sebagian otak tidak mendapatkan pasokan darah yang membawa oksigen sehingga sel/jaringan mengalami kematian [2].

Jumlah pasien stroke meningkat setiap tahunnya, ini akan berdampak buruk terhadap kesehatan, maka untuk meminimisir dampak buruk dari penyakit stroke maka diperlukan kesadaran berbagai tanda peringatan stroke dapat mencegah sedini mungkin terhadap serangan stroke [4]. Pencegahan lain untuk mengetahui peringatan stroke tahap awal dengan merancang sebuah pendekatan yang dapat mengidentifikasi dan memprediksi apakah seorang dapat terkena stroke atau tidak [5] dengan mempertimbangkan faktor risiko utama dari stroke seperti hipertensi, riwayat penyakit jantung, merokok, serta faktor gaya hidup seperti tingkat obesitas, aktivitas fisik, tempat tinggal dll [6]. Machine learning dalam hal memainkan peranan penting dalam identifikasi dan prediksi [7] dari data-data informasi dan kejadian masa lalu yang diketahui [8] dalam diagnosis tahap awal, agar dapat mengurangi risiko kemungkinan terkena stroke dengan mengambil tindakan yang diperlukan.

Penelitian-penelitian sebelumnya menunjukkan perkembangan prediksi algoritma machine learning pada penyakit stroke seperti Weighted Voting [4], Gradient Boosting algorithm [5], decision tree [9] [10] [11], Artificial Neural Network (ANN) [12] [13], random forest [9] [8] [7], logistic regression [7], Naïve Bayes Classification [14] [9] menggunakan data dari kaggle stroke prediction, dan deep neural network pada penelitian [15] beberapa penelitian seperti [4], [5] dan [7] sudah mencapai hasil baik. Namun, beberapa algoritma seperti logistik regresssion pada penelitian [4], [5], random forest pada penelitian [5] dan Decision tree pada penelitian [4], [5] masih mencapai akurasi yang kurang, ini disebabkan karena ketidakseimbangan data, tantangannya adalah bagaimana dapat menyeimbangkan kelas untuk mendapatkan hasil algortima yang terbaik.

Pada penelitian ini menganalisis 5110 data historis pasien yang terkena stroke dan tidak, dengan mengkorelasikan data-data gaya hidup, faktor risiko medis dan faktor yang tidak dapat dikenadalikan sebagai data



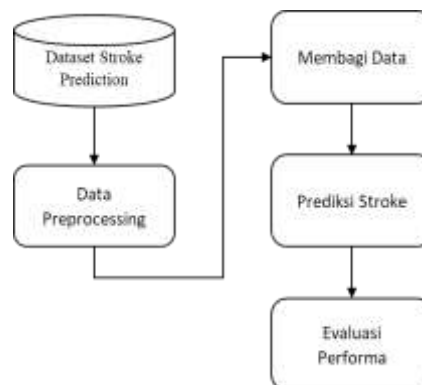
inputan/feature dan pada penelitian menganalisis performa algoritma seperti Multi layer perceptron, KNN, Decision Tree dan Random Forest dengan menerapkan Synthetic Minority Oversampling Technique (SMOTE) untuk penanganan imbalanced class pada data.

2. METODOLOGI PENELITIAN

2.1 Description Material

Dalam penelitian ini, data yang digunakan adalah dataset dari kaggle [16] yang di dalamnya terdapat informasi klinis di dapatkan dari Klinik Medis Bangladesh, dataset terdiri dari 5110 entri pasien dengan 12 atribut. Dalam data tersebut 11 feature pertama adalah pertanyaan yang digunakan untuk dapat memprediksi apakah pasien tersebut bisa terkena stroke atau tidak, pertanyaan tersebut dapat dibagi menjadi 3 bagian yaitu faktor gaya hidup, faktor risiko medis, dan faktor yang tidak dapat dikendalikan [5]. Faktor gaya hidup seperti kebiasaan makan, kebiasaan minum, aktivitas fisik dan merokok. Faktor risiko medis yaitu kemungkinan stroke dapat dikendalikan seperti tekanan darah, level glukosa, dan pasien pernah terkena serangan jantung. Sedangkan yang tidak dapat dikendalikan seperti jenis kelamin, usia dan etnis. Deskripsi ke 11 fitur tersebut adalah

1. **id**: pengenal unik
2. **gender**: "Pria", "Wanita" atau "Lainnya"
3. **age**: usia pasien
4. **hypertension**: 0 jika pasien tidak memiliki hipertensi, 1 jika pasien memiliki hipertensi
5. **heart_disease**: 0 jika pasien tidak memiliki penyakit jantung, 1 jika pasien memiliki penyakit jantung
6. **ever_married**: "Tidak" atau "Ya"
7. **work_type**: "anak-anak", "Govt_jov", "Tidak pernah bekerja", "Swasta" atau "Wiraswasta"
8. **Residence_type**: "Pedesaan" atau "Perkotaan"
9. **avg_glucose_level**: tingkat glukosa rata-rata dalam darah
10. **bmi**: indeks massa tubuh
11. **smoking_status**: "sebelumnya merokok", "tidak pernah merokok", "merokok" atau "Tidak diketahui"*
12. **stroke**: 1 jika pasien mengalami stroke atau 0 jika tidak



Gambar 1. Kerangka Kerja model evaluasi prediksi stroke

2.2 Kerangka Kerja

Kerangka kerja yang diusulkan untuk melatih algoritma dapat memprediksi apakah seseorang dapat terkena stroke atau tidak dimulai dari (1) pengoleksian data, (2) Preprocessing data (3) membagi data menjadi data training dan testing (4) melatih algoritma klasifikasi, dan (6) evaluasi algoritma klasifikasi. Gambar 1 menunjukkan kerangka kerja prediksi stroke dengan algoritma pengklasifikasi.

2.3 Data Preprocessing

Setelah pengumpulan data yang sesuai, langkah selanjutnya adalah menyiapkan dataset agar algoritma pengklasifikasi dapat berkekerja dengan baik dan efisien langkah ini disebut dengan preprocessing data. Langkah ini sangat penting dilakukan, agar algoritma pengklasifikasi dapat menghasilkan nilai terbaiknya. Yang dilakukan pada preprocessing data adalah menghilangkan noise dan outliere pada data, menghilangkan missing value atau data yang hilang, menghapus data duplikat, pengkodean, dan penanganan data yang tidak seimbang menggunakan SMOTE.

2.4 Membagi data (Splitting Data)

Setelah preprocessing data selesai, langkah selanjutnya adalah Splitting data atau membagi data berarti membagi data menjadi data training dan data testing. Pada penelitian ini data dibagi menjadi 80% data training dan 20% data testing dan menguji dengan 10K-Fold Validation



2.5 Algoritma Kalasifikasi

Setelah data dibagi menjadi data training dan data testing, langkah selanjutnya melatih algoritma pengklasifikasi. Dalam penelitian ini mengadopsi algoritma pengklasifikasi seperti Multi layer perceptron, KNN, Decission Tree dan Random Forest. Semua algoritma pengklasifikasi di latih dan di uji menggunakan data set testing sebesar 20% dan evaluasi hasilnya.

2.6 Evaluasi

Langkah selanjutnya yaitu mengevaluasi hasil algoritma pengklasifikasi, dalam penelitian ini teknik evaluasi model algoritma pengklasifikasi yang digunakan adalah confusion matrix. Confusion matrix adalah cara termudah untuk mengevaluasi hasil kinerja algoritma pengklasifikasi dengan membandingkan beberapa contoh positif yang diklasifikasikan benar atau salah, dan negatif yang diklasifikasikan benar atau salah. Confusion matrix mempunyai pandangan beragam, dimana confusion matrix mempunyai peran mendasar dalam mengevaluasi kinerja algoritma klasifikasi [17]. Dalam confusion matrix, baris merepresentasikan label yang benar, dan kolom merepresentasikan hasil prediksi algoritma pengklasifikasi, **Tabel 1** menunjukkan confusion matrix performance.

Tabel 1. Confusion Matrix

	Prediksi Stroke	Prediksi Tidak Stroke
Stroke	TP	TN
Tidak Stroke	FP	FN

TP (True Positif) merupakan data positif stroke yang di prediksi benar oleh algoritma pengklasifikasi.

TN (True Negatif) merupakan data positif stroke yang di prediksi salah oleh algoritma pengklasifikasi.

FP (False Positif) merupakan data tidak terkena stroke dan algoritma pengklasifikasi memprediksi dengan benar.

FN (False Negatif) merupakan data tidak terkena stroke dan algoritma pengklasifikasi memprediksi dengan salah.

Accuracy dapat menggambarkan seberapa akurat model algoritma pengklasifikasi dapat memprediksi dengan benar, rumus untuk menghitung akurasi adalah

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk menganalisis dan membuat sebuah model algoritma pengklasifikasi untuk bisa memprediksi stroke, dari data-data histories pasien yang telah terkena dan yang tidak terkena. 7 algoritma pengklasifikasi seperti Multi layer perceptron, KNN, Decission Tree dan Random Forest diusulkan dan di analisis hasil performanya dengan menggunakan metode confusion matrix dan dihitung nilai akurasi. menggunakan bahasa pemrograman python dengan menggunakan google colab dan environment seperti jupyter notebook dan library seperti numpy, pandas, matplotlib, tensorflow dan sklearn.

Tabel 2. Confusion matrix hasil Komparasi algoritma pengklasifikasi

Algoritma Klasifikasi	Label	Prediksi	
		Stroke	No. Stroke
MLP	Stroke	11	0
	No. Stroke	0	703
KNN	Stroke	11	0
	No. Stroke	702	1
Decission Tree	Stroke	11	0
	No. Stroke	0	703
Random Forest	Stroke	702	1
	No. Stroke	11	0

Tabel 2 adalah confusion matrix dari algoritma pengklasifikasi yang dilakukan training untuk mengukur performa prediksi stroke

Tabel 3. Komparasi Hasil Akurasi Algoritma

	Accuracy 80%	10K-Fold Validation
MLP	100	99,94
KNN	99,9	99,82
Decission Tree	100	99,86
Random Forest	99,99	99,99



Error! Reference source not found. adalah hasil akurasi dari 4 algoritma pengklasifikasi hasil dari training dan testing. Nilai dari tabel dapat kita simpulkan bahwa terdapat 2 perbedaan pengujian, yaitu split data 80% untuk train dan 20% untuk testing dan 10K Fold Validation, keduanya mempunyai hasil yang berbeda, namun dapat ditarik kesimpulan bahwa algoritma yang diusulkan dapat bekerja dengan baik pada prediksi stroke, seperti pada random forest dengan menggunakan split validation, hasil akurasinya sebesar 99,99% atau bisa dikatakan sempurna, dan menggunakan 10K validation mendapatkan hasil akurasi terbaik yaitu 99,95%. Dibuktikan dengan hasil pada python pada **Gambar 2**

```
Random Forest :
[0.99821109 1.          1.          1.          0.99821109 1.
 1.          0.99821109 1.          1.          ]
Accuracy: 0.9985994397759104
K-Foldcv mean accuracy: 99.9463 %
```

Gambar 2. Hasil Accuracy Random Forest

4. KESIMPULAN

Dalam penelitian ini, bertujuan untuk membuat sebuah model yang dapat memprediksi stroke dari dataset publik dari kaggle, ini bertujuan untuk mendeteksi dini apakah seseorang bisa terkena stroke atau tidak, dari hasil analisis dapat disimpulkan bahwa model yang diusulkan seperti Multi Layer Perceptron, KNN, Decision Tree dan Random Forest dapat bekerja dengan baik dengan hasil akurasi yang memuaskan yaitu hampir 100%. Penelitian selanjutnya dapat dikembangkan sebuah aplikasi yang dapat memprediksi stroke berbasis web atau mobile.

REFERENCES

- [1] M. P. Lindsay *et al.*, "Global Stroke Fact Sheet 2019," 2019.
- [2] Kemenkes RI, "Stroke Dont Be The One." InfoDATIN, Jakarta, p. 10, 2019.
- [3] T. Truelsen, S. Begg, and C. Mathers, "The global burden of cerebrovascular disease," *Glob. Burd. Dis.*, pp. 1–67, 2016.
- [4] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. Al Mamun, and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2020*, pp. 1464–1469, 2020.
- [5] M. Rajora, M. Rathod, and N. S. Naik, "Stroke Prediction Using Machine Learning in a Distributed Environment," *Proc. 17th Int. Conf. ICDCIT 2021 Bhubaneswar, India, January 7–10, 2021*, pp. 238–266, 2021.
- [6] A. Guzik and C. Bushnell, "Stroke epidemiology and risk factor management," *Contin. Lifelong Learn. Neurol.*, vol. 23, no. 1, pp. 15–39, 2017.
- [7] M. S. Azam, Habibullah, and H. K. Rana, "Performance Analysis of Various Machine Learning Approaches in Stroke Prediction," *Int. J. Comput. Appl.*, vol. 175, pp. 11–15, 2020.
- [8] J. Alberto and T. Rodríguez, "Stroke prediction through Data Science and Machine Learning Algorithms," no. M1, 2021.
- [9] D. Babu, V. Karunakaran, S. Gopinath, S. P. Angeline Kirubha, S. Latha, and P. Muthu, "Gui based prediction of heart stroke using artificial intelligence," *Mater. Today Proc.*, vol. 47, pp. 104–108, 2021.
- [10] R. S. Rohman, R. A. Saputra, and D. A. Firmansaha, "Komparasi Algoritma C4.5 Berbasis PSO Dan GA Untuk Diagnosa Penyakit Stroke," *CESS (Journal Comput. Eng. Syst. Sci.)*, vol. 5, no. 1, p. 155, 2020.
- [11] N. D. Saputri, "Komparasi penerapan metode Bagging dan Adaboost pada Algoritma c4. 5 untuk prediksi Penyakit Stroke," UIN Sunan Ampel Surabaya, 2021.
- [12] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke Prediction Using Smote-Tomek and Neural Network," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–5.
- [13] L. Menezes, E. Gnanaraj, S. Bindra, and A. Pansare, "Early Stage Stroke Prediction Using Artificial Neural Network," in *2021 2nd Global Conference for Advancement in Technology (GCAT)*, 2021, pp. 1–6.
- [14] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 539–545, 2021.
- [15] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, p. 101723, 2019.
- [16] Fedesoriano, "Stroke Prediction Dataset," *Kaggle*, 2021. [Online]. Available: <https://www.kaggle.com/fedoriano/stroke-prediction-dataset>.
- [17] J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," *Inf. Sci. (Ny.)*, vol. 507, pp. 772–794, 2020.