Jurnal Sistem Komputer dan Kecerdasan Buatan
Vol.VI No.1 September Tahun 2022

Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke

Felix Indra Kurniadi[1], Pramitha Dwi Larasati[2]

[1]Computer Science Department, School of Computer Science, Bina Nusantara University
[2]School of Engineering and Technology, Information System, Tanri Abeng University
Felix.indra@binus.ac.id, pramitha.dwi@tau.ac.id

Diterima : 22 Agustus 2022
Disetujui : 01 Oktober 2022

***Abstract*—Stroke merupakan salah satu penyakit yang berbahaya di dunia penyakit stroke merupakan penyakit kedua yang mengakibatkan kematian. Pada saat ini proses pendeteksian factor resiko seseorang untuk terkena stroke sangat penting dilakukan sebagai early detection. Pada saat ini sudah banyak algoritma machine learning yang mencoba mengatasi permasalahan dalam clinical data seperti SVM, dan Random Forest. Kedua metode ini sayangnya memiliki problem utama terbesar yaitu mudah sekali overfitting dan sangat rentan terhadap noise. Disebabkan oleh kelemahan yang diusulkan oleh kedua metode ini, peneliti mengusulkan metode Light Gradient Boosting Machine. Light Gradient Boosting Machine merupakan algoritma yang memiliki computational cost rendah. Pada penelitian ini kita menggunakan dua scenario utama yaitu scenario tanpa menggunakan fitur seleksi dan scenario kedua dengan menggunakan fitur seleksi menggunakan Variance Threshold method. Kesimpulan yang didapatkan dari penelitian adalah metode Light GBM memiliki hasil yang seimbang antara SVM dan RF akan tetapi model yang dibuat sangat bias hal ini dapat dilihat dari nilai precision dan recall yang berbeda jauh dari nilai akurasinya.**

***Keywords* —light GBM, SVM, Random Forest, Stroke**

I. PENDAHULUAN

Stroke merupakan salah satu penyakit yang menyerang otak. Kondisi ini dikarenakan pasokan darah ke otak terganggu disebabkan oleh penyumbatan di dalam otak atau pecahnya pembuluh darah di dalam otak. Akibat dari gangguan *stroke* ini biasanya mempengaruhi kemampuan manusia dalam mengendalikan area tubuh yang dikontrol oleh otak [1].

Stoke sendiri memiliki beberapa gejala seperti lemah pada wajah yang menyebabkan wajah tidak berbentuk simetris, kesulitan dalam berbicara, kesemutan, keterbatasan dalam pergerakan seperti kesulitan dalam melakukan aktifitas yang membutuhkan otot dalam tubuh seperti mengangkat kedua lengan dan lainnya [1]. Gejala tersebut menyebabkan kesulitan pasien dalam menjalankan kehidupannya sehari-hari.

Berdasarkan data yang didapatkan dari *World Stroke Organization*, tiap tahunnya terdapat 13,7 juta kasus baru dan 5,5 juta kematian yang diakibat penyakit ini. Data juga menyebutkan bahwa sekitar 70% penderita *stroke* merupakan orang-orang yang berada pada negara dengan pendapatan rendah dan menengah [2].

Selain berdampak terhadap Kesehatan personal, *stroke* di Indonesia juga membebani negara dalam pembiayaan Kesehatan. Berdasarkan data yang dihimpun dari Badan Penyelenggara Jaminan Sosial (BPJS) Kesehatan dimana total yang dihabiskan oleh penyakit stroke di Indonesia adalah 2,56 triliun rupiah pada tahun 2018. Nilai ini akan terus bertambah seiring dengan bertambahnya penderita *stroke* di Indonesia [2].

Berdasarkan problematika ini diperlukan sebuah sistem yang dapat melakukan deteksi penyakit *stroke*. Beberapa penelitian mencoba melakukan pendeteksian terhadap *Ishemic Stroke*

dengan pendekatan computer vision seperti yang dilakukan oleh [3]–[6]. Beberapa penelitian lainnya mencoba melakukan pendeteksian atau prediksi menggunakan clinical data seperti yang dilakukan oleh [7], [8].

Perkembangan jaman membuat data terutama *clinical data* sangat mudah didapatkan. Banyaknya informasi data mentah dalam dunia Kesehatan dapat diekstraksi dan dibuat klasifikasi menggunakan metode *machine learning* saat ini [8]. Beberapa metode *machine learning* yang digunakan seperti *Support Vector Machine* (SVM) yang dilakukan oleh [8] dan *Random Forest* yang dilakukan oleh [7].

Random Forest dan SVM merupakan salah satu metode machine learning yang sering digunakan dalam penyelesaian masalah supervised dalam machine learning akan tetapi kedua metode yang disebutkan sebelumnya memiliki kelemahan. SVM merupakan algoritma yang berfokus pada menentukan *hyperplane* untuk memisahkan data. Data yang memiliki banyak derau akan sangat mempengaruhi performa dari model ini, dilain pihak dengan menggunakan data yang banyak akan mempengaruhi kecepatan dari computational time dari model [9].

Berbeda dengan SVM, *Random Forest* bekerja baik terhadap data yang memiliki derau. Sayangnya, problem yang muncul dalam *Random Forest* terdapat dalam *complexity* hal ini disebabkan karena *Random Forest* menciptakan banyak *tree* sehingga membutuhkan banyak computational power. Hal ini juga mempengaruhi *training time* yang dibutuhkan dalam menyelesaikan permasalahan ini [10]. Selain itu sulit untuk melakukan interpretasi terhadap hasil dari *Random Forest* [11].

Pada tahun 2017, Ke et al. mengusulkan sebuah algoritma yang disebut dengan *Light Gradient Boosting Machine* (Light GBM) [12]. Keuntungan dari metode ini adalah memiliki training yang cepat dan lebih efisien, membutuhkan memori yang lebih rendah dalam penggunaannya, menghasilkan akurasi yang lebih baik dan mampu mengatasi data yang besar [12].

Berdasarkan dari keuntungan *Light GBM* diatas maka penulis ingin mengimplementasikan metode *Light GBM* dalam melakukan klasifikasi *Ischemic Stroke* dari *clinical data*.

Artikel ilmiah ini akan terdiri dari lima bab. Bab I membahas mengenai pendahuluan mengenai penelitian yang dilakukan dan menjelaskan mengenai penelitian yang sudah dilakukan oleh penulis lainnya. Bab II akan menjelaskan mengenai teori dari *Light GBM*, Bab III akan membahas mengenai metodologi yang digunakan dalam penelitian ini. Bab IV menjelaskan mengenai Eksperimen dan Pembahasan dari eksperimen yang dilakukan. Bab V akan membahas mengenai kesimpulan dan penelitian yang akan dilakukan selanjutnya.

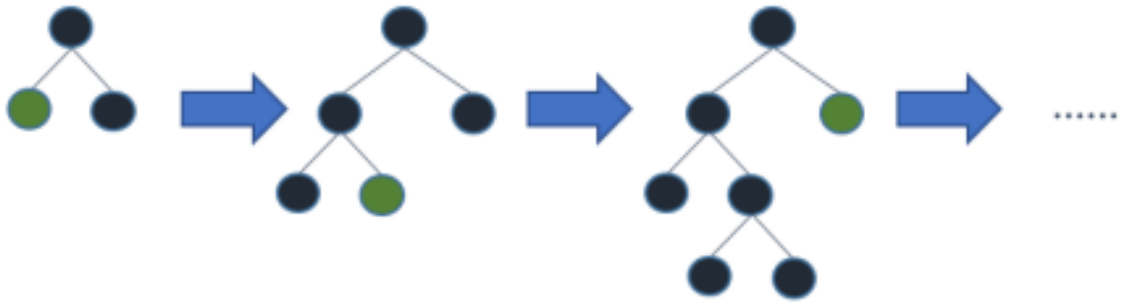
II. LIGHT GRADIENT BOOSTING MACHINE

Light Gradient Boosting Machine (Light GBM) merupakan sebuah metode *gradient boosting* yang cepat, terdistribusi dan memiliki *high-performance* berbasiskan *decision tree*. *Light GBM* merupakan salah satu metode *ensemble* yang melakukan agrerasi terhadap prediksi dari beberapa *decision tree* (dengan menambahkan setiap *tree*)[12].

Asumsikan kita akan membuat sebuah model *Light GBM* dengan *tree* (T). Menggunakan konsep *additive training process* untuk sebuah dataset kita dapat memformulasikan rumusan prediksi[12], [13]

$$\hat{y}_i^{(t)} = \sum_{i=1}^{n} f_n(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (1)$$

Light GBM merupakan implementasi *Gradient Boosting Decision Tree (GBDT)*. Pada proses pelatihan setiap individual *decision tree* (f) akan melakukan pemisahan data. Light GBM menggunakan dua strategi yaitu *gradient-based one-side sampling (GOSS)* dan *leaf-wise growth*[13]. Konsep *leaf-wise growth* akan digambarkan pada Gambar 1.



Gambar 1. Leaf wise growth[13]

Konsep *leaf-wise growth* merupakan salah satu Teknik untuk membatasi depth dari model *Light GBM*, proses ini dilakukan untuk mencari *node* dengan *splitting gains* terbesar. Proses yang dilakukan selanjutnya adalah memecahkan *node* tersebut dan meneruskan untuk *node* yang baru. Model *Light GBM* tidak perlu menambahkan kedalaman model untuk menghindari penggunaan daya komputasi yang lebih besar dan juga mengurangi *overfitting* [12].

III. METODOLOGI

Pada bab ini akan menjelaskan mengenai metodologi yang digunakan dalam melakukan proses penelitian. Pada Gambar 2 akan menjelaskan mengenai metodologi yang kami gunakan untuk pembuatan penelitian ini.

Gambar 2. Metodologi Penelitian

A. Pengambilan Data

Dataset yang digunakan dalam penelitian ini diambil dari data *The Electronic Health Record* yang dikontrol oleh *McKinsey & Company*. Data ini merupakan salah satu bagian dari *hackaton* yang dilakukan oleh perusahaan *McKinsey & Company*. Dataset terdiri atas data

pasien yang berjumlah 29,072 orang dengan 12 attribute yang digunakan. 11 attribute akan digunakan sebagai variable bebas yaitu *age, gender, marital status, patient identifier, work type, residence type, heart condition, glucose level dan hypertension*.

Tabel 1. Distribusi Label

	Jumlah
Stroke	548
Non-stroke	28524

B. Data Transformasi

Beberapa Langkah untuk mempermudah proses pengektraksian adalah dengan menggunakan apa yang dilakukan oleh [8]. Beberapa fitur seperti *glucose level, BMI* dan *Age* mengalami proses diskritasi. Secara medis data glucose level dibagi menjadi 4 tipe berbeda yaitu *hypoglycemia*, normal, *prediabetes* dan *diabetes*. Hal yang sama kita lakukan dengan data BMI dimana akan dibagi menjadi empat yaitu, *underweight, normal, overweight* dan obesitas. Kita membagi juga fitur age menjadi 4 bagian yaitu anak-anak, remaja, dewasa, dan lansia [8]. Pada artikel ini kami tidak melakukan proses pengisian missing value, Data yang memiliki missing value akan kami hapus sehingga tidak perlu melakukan proses pencarian nilai missing value.

C. Seleksi Fitur

Pada penelitian ini kami menggunakan metode seleksi fitur untuk menyelesaikan permasalahan pada banyaknya attribute. Kami menggunakan metode *Variance Threshold*. Konsep *Variance Threshold* adalah membuah fitur yang memiliki nilai dibawah threshold nya. Pada eksperimen ini kami akan menggunakan threshold 0.8

D. Modelling

Proses modelling pada metode ini menggunakan metode *Light GBM* yang sudah dijelaskan pada Bab II. *Light GBM* ini menjadi acuan utama dan untuk membuktikan bahwa metode *Light GBM* merupakan metode yang baik digunakan dalam artikel ilmiah ini kami membandingkan metode *Light GBM* dengan beberapa metode lainnya seperti *Support Vector Machine* dan *Random Forest*[7], [8].

E. Evaluasi

Evaluasi yang digunakan pada penelitian ini adalah *accuracy, precision,* dan *recall.*

IV. EKSPERIMEN

Pada bab eksperimen, kami akan menjelaskan mengenai pengaturan eksperimen yang kami gunakan dan hasil dari eksperimen yang kami lakukan.

A. Pengaturan Eksperimen

Pada penelitian eksperimen, kami menggunakan google collab dalam pembuatannya dan mendapatkan bantuan dari beberapa library machine learning seperti scikit-learn dan pandas. Tabel 2 menjelaskan mengenai hyperparameter yang digunakan dalam pembuatan artikel ini.

Tabel 2. Hyperparameter Metode

Method	Hyperparameter
Light GBM	num_leaves = 31 learning_rate = 0.1 n_estimators = 100 min_child_samples = 20 subsample = 1.0 reg_alpha = 0.0
SVM	C=1.0 Tolerance = 1e-4 Loss = squared_hinge
Random Forest	n_estimators = 100 criterion = gini min_sample_split = 2 min_sample_leaf=1 max_features = sqrt

Hyperparameter yang ditampilkan pada Table 1 merupakan default hyperparameter yang disediakan oleh scikit-learn. Kami menggunakan ini untuk meminimalisir pengaruh dari hyperparameter changing yang akan terjadi pada penelitian ini.

B. Hasil Penelitian

Penelitian ini menggunakan dua scenario yang akan dicobakan yaitu scenario satu merupakan scenario yang membahas mengenai hasil yang didapatkan tanpa menggunakan proses fitur seleksi dan scenario kedua adalah merupakan hasil yang didapatkan setelah mendapatkan hasil dari fitur seleksi.

Tabel 3 dan Tabel 4 memberikan gambaran hasil yang didapatkan dari scenario 1 dan scenario 2.

Tabel 3 Hasil dari precision scenario 1

Method	accuracy	precision	recall
Light GBM	98	0.5	0.49
SVM	98	0.5	0.49
Random Forest	98	0.51	0.56

Tabel 4. Hasil dari precision scenario 2

Method	accuracy	precision	recall
Light GBM	98	0.5	0.49
SVM	98	0.5	0.49
Random Forest	98	0.5	0.49

Berdasarkan hasil yang ditampilkan pada Tabel 3 dan Tabel 4 dapat dikatakan bahwa hasil akurasi memiliki nilai yang baik dimana hasil nilai akurasi mencapai 98% akan tetapi jika kita melihat dari nilai *precision* dan *recall* kedua tabel diatas dapat disimpulkan bahwa model masih memiliki bias terhadap data yang ada saat ini.

Pada hasil yang diberikan di scenario kedua dapat dilihat bahwa tidak ada perbedaan antara ketiga metode. Berdasarkan hasil yang didapatkan dapat disimpulkan bahwa penggunaan feature selection Variance Threshold tidak memperbaiki model tetapi hanya menurunkan nilai dari precision dan recall.

Beberapa kemungkinan terjadinya bias akan tetapi jika kita melihat kepada data yang dimiliki saat ini adalah imbalanced. Dimana data penderita stroke lebih sedikit dibandingkan dengan data hanya sebesar 548.

V. SIMPULAN

Stroke merupakan salah satu penyakit yang mematikan di dunia. Kebutuhan terhadap sistem yang dapat memberikan pemberitahuan resiko terhadap penyakit stroke sangat dibutuhkan saat ini. Metode yang diusulkan pada penelitian ini adalah *Light Gradient Boosting Machine* (Light GBM) dengan dibandingkan dengan beberapa metode machine learning yang terkenal seperti SVM dan *Random Forest*.

Berdasarkan hasil dari eksperimen yang dilakukan pada penelitian ini dapat diambil beberapa kesimpulan. Pada data yang dimiliki proses fitur seleksi diusulkan tidak memberikan pengaruh terlalu berbeda hal ini berbeda dengan statement yang diberikan oleh [8] pada artikel ilmiahnya yang mengatakan proses fitur seleksi dibutuhkan. Hasil dari akurasi diberikan dari ketiga metode memiliki nilai yang sama di kedua scenario akan tetapi hasil *precision* dan *recall* pada scenario memberikan hasil yang berbeda pada metode *Random Forest.* Nilai precision dan recall sangat jauh berbeda dengan nilai akurasi ini mengindikasikan bahwa adanya imbalanced data yang tidak ditangani akan membuat bias.

Pada penelitian selanjutnya terhadap data stroke dibutuhkan sebuah penyelesaian dalam penentuan model hyperparameter atau mengatasi persoalan bias yang terjadi pada data yang saat ini digunakan. Salah satu cara yang diusulkan untuk persoalan ini adalah dengan menerapkan metode seperti undersampling atau oversampling untuk mengatasi permasalahan yang ada. Selain sampling dibutuhkan juga penelitian dibutuhkan

juga pencarian hyperparameter terbaik dengan menggunakan metode-metode pencarian hyperparameter.

DAFTAR PUSTAKA

[1] alodokter, "Stroke," *Alodokter*. https://www.alodokter.com/stroke (accessed Aug. 22, 2022).

[2] Kementrian Kesehatan RI, "Stroke Don't be the one." Pusat Data dan Informasi Kementrian Kesehatan RI. Accessed: Aug. 22, 2022. [Online]. Available: https://pusdatin.kemkes.go.id/download.php?file=download/pusdatin/infodatin/infodatin-stroke-dont-be-the-one.pdf

[3] S. Joshi and S. Gore, "Ishemic Stroke Lesion Segmentation by Analyzing MRI Images Using Dilated and Transposed Convolutions in Convolutional Neural Networks," in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, Aug. 2018, pp. 1–5. doi: 10.1109/ICCUBEA.2018.8697545.

[4] S. Gupta, A. Mishra, and Menaka R, "Ischemic Stroke detection using Image processing and ANN," in *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, India, May 2014, pp. 1416–1420. doi: 10.1109/ICACCCT.2014.7019334.

[5] A. F. Z. Yahiaoui and A. Bessaid, "Segmentation of ischemic stroke area from CT brain images," in *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, Tunis, Tunisia, 2016, pp. 13–17. doi: 10.1109/ISIVC.2016.7893954.

[6] F. Aboudi, C. Drissi, and T. Kraiem, "Efficient U-Net CNN with Data Augmentation for MRI Ischemic Stroke Brain Segmentation," in *2022 8th International Conference on Control, Decision and Information Technologies (CoDIT)*, Istanbul, Turkey, May 2022, pp. 724–728. doi: 10.1109/CoDIT55151.2022.9804030.

[7] A. Vrtkova, "Predicting clinical status of patients after an acute ischemic stroke using random forests," in *2017 International Conference on Information and Digital Technologies (IDT)*, Zilina, Slovakia, Jul. 2017, pp. 417–422. doi: 10.1109/DT.2017.8024330.

[8] M. S. Pathan, Z. Jianbiao, D. John, A. Nag, and S. Dev, "Identifying Stroke Indicators Using Rough Sets," *IEEE Access*, vol. 8, pp. 210318–210327, 2020, doi: 10.1109/ACCESS.2020.3039439.

[9] K. Monien and R. Decker, "Strengths and Weaknesses of Support Vector Machines Within Marketing Data Analysis," in *Innovations in Classification, Data Science, and Information Systems*, D. Baier and K.-D. Wernecke, Eds. Berlin/Heidelberg: Springer-Verlag, 2005, pp. 355–362. doi: 10.1007/3-540-26981-9_41.

[10] K. Huo and N. Singh, *Ace the data science interview: 201 real interview questions asked by FAANG, tech startups, & Wall Street*. United States: Ace the Data Science Interview, 2021.

[11] M. Aria, C. Cuccurullo, and A. Gnasso, "A comparison among interpretative proposals for Random Forests," *Mach. Learn. Appl.*, vol. 6, p. 100094, Dec. 2021, doi: 10.1016/j.mlwa.2021.100094.

[12] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," Dec. 2017. [Online]. Available: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

[13] T. Chen *et al.*, "Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine," *IEEE Access*, vol. 7, pp. 150960–150968, 2019, doi: 10.1109/ACCESS.2019.2946980.