

PREDIKSI HARGA SAHAM DI BURSA EFEK INDONESIA MENGUNAKAN ALGORITMA RANDOM FOREST DAN LIGHTGBM

Syarip Chumaeni¹, Nana Suarna², Agus Bahtiar³, Rosidin⁴.

Program Studi Teknik Informatika¹²
Program Studi Sistem Informasi³⁴

STMIK IKMI Cirebon
<https://ikmi.ac.id/page/18/?lang=de>
syarifchumaeni2021@gmail.com

(*) Corresponding Author : syarifchumaeni2021@gmail.com
Published : 30 Juni 2026

Abstract—The capital market in Indonesia is highly dynamic and complex, so predicting stock prices requires an analytical approach capable of capturing nonlinear patterns and high variability. With advancements in machine learning, there is significant potential to improve prediction accuracy through advanced algorithms. This study aims to analyze and compare the performance of two ensemble learning algorithms—Random Forest and Light Gradient Boosting Machine (LightGBM)—in predicting stock prices based on company fundamental variables such as market capitalization, number of shares outstanding, industry sector, and listing board. The secondary data I used comes from the Indonesia Stock Exchange (IDX) stock dataset, comprising a total of 829 entries and 14 attributes. I processed this data through data cleaning, encoding of categorical variables, feature normalization, data splitting (80% training, 20% testing), and modeling using both algorithms, namely: Random Forest with an RMSE of 2949.262580, MAE of 1675.096152, and R^2 of -2.122673, meaning this model still requires improvement but is relatively more stable than LightGBM. Meanwhile, LightGBM achieved an RMSE of 2293.665978, an MAE of 1628.837321, and an R^2 of -0.888687. From the feature importance analysis, it turns out that Shares, ListingBoard, and Sector are the most influential features for stock price prediction, in line with asset valuation theory, which emphasizes the role of company characteristics in determining market value. This study demonstrates that machine learning methods, particularly Random Forest, can serve as a useful predictive approach for investors and capital market analysts in Indonesia. These findings contribute to the development of data-driven investment decision support systems. However, there are limitations, such as the exclusion of macroeconomic variables and the lack of time-series data. For future research, I suggest integrating external variables, applying cross-validation techniques, and exploring other algorithms such as XGBoost, CatBoost, or deep learning models.

Keywords: Stock price prediction, Random Forest, LightGBM, Machine Learning, Indonesia Stock Exchange, Ensemble Learning.

Abstrak—Pasar modal di Indonesia itu sangat dinamis dan kompleks, jadi untuk memprediksi harga saham, butuh pendekatan analitis yang bisa menangkap pola *non-linear* dan variabilitas tinggi. Dengan perkembangan *machine learning*, ada peluang besar untuk meningkatkan akurasi prediksi lewat algoritma canggih. Penelitian saya ini bertujuan menganalisis dan membandingkan performa dua algoritma ensemble learning, yakni Random Forest dan *Light Gradient Boosting Machine* (LightGBM), dalam memprediksi harga saham berdasarkan variabel fundamental perusahaan seperti kapitalisasi pasar, jumlah saham beredar, sektor industri, dan papan pencatatan. Data sekunder yang saya gunakan berasal dari *dataset* saham Bursa Efek Indonesia (BEI) dengan total 829 entri dan 14 atribut, yang kemudian saya proses melalui pembersihan data, *encoding* variabel kategorikal, normalisasi fitur, pembagian data (80% pelatihan, 20% pengujian), serta pemodelan menggunakan kedua algoritma yaitu. Random Forest dengan nilai RMSE sebesar 2949.262580, MAE 1675.096152, dan R^2 -2.122673, yang artinya model ini masih perlu perbaikan tapi relatif lebih stabil ketimbang *LightGBM*. Sedangkan LightGBM mendapat RMSE 2293.665978, MAE 1628.837321 dan R^2 -0.888687, Dari analisis *feature importance*, ternyata Shares, ListingBoard, dan Sector adalah fitur paling berpengaruh terhadap prediksi harga saham, sesuai dengan teori valuasi aset yang menekankan peran karakteristik perusahaan dalam menentukan nilai pasar. Penelitian ini membuktikan bahwa metode *machine learning*, terutama Random Forest, bisa jadi pendekatan prediktif yang berguna buat investor dan analis pasar modal di Indonesia. Temuan ini berkontribusi dalam pengembangan sistem

pendukung keputusan investasi berbasis data. Meski begitu, ada keterbatasan, seperti tidak memasukkan variabel makroekonomi dan tidak menggunakan data *time series*. Untuk penelitian selanjutnya, saya sarankan untuk mengintegrasikan variabel eksternal, menerapkan teknik validasi silang, dan mengeksplorasi algoritma lain seperti *XGBoost*, *CatBoost*, atau model *deep learning*.

Kata Kunci: Prediksi harga saham, Random Forest, LightGBM, *Machine Learning*, Bursa Efek Indonesia, *Ensemble Learning*.

INTRODUCTION

Perkembangan pesat di bidang informatika telah mengubah berbagai aspek kehidupan, dari teknologi dan bisnis hingga pendidikan, sehingga data menjadi aset paling berharga di era digital saat ini. Kemajuan dalam ilmu komputer, terutama di domain kecerdasan buatan dan pembelajaran mesin, membuka peluang baru untuk menyelesaikan masalah kompleks yang sulit diatasi dengan metode tradisional. Di konteks finansial, pasar modal adalah salah satu area yang paling diuntungkan oleh inovasi informatika, karena menyediakan data dalam volume besar dan kecepatan tinggi. Analisis data *time series* finansial kini membutuhkan algoritma yang mampu menangkap pola *non-linear* dan dinamika pasar yang sangat volatil. Oleh karena itu, penerapan teknik-teknik canggih dari informatika sangat penting untuk meningkatkan efisiensi dan akurasi dalam pengambilan keputusan investasi.

Fluktuasi harga saham di pasar negara berkembang, seperti Indonesia, menampilkan tantangan kompleks akibat interaksi variabel makroekonomi, arus modal global, dan sentimen pasar yang cepat berubah. Studi menunjukkan bahwa ketidakpastian global dan faktor seperti krisis kesehatan dapat memperburuk volatilitas pasar saham di negara-negara *emerging market* [1]. Volatilitas ini tidak hanya memengaruhi harga aset, tetapi juga secara signifikan memengaruhi perilaku investor individu. Sebagai contoh, dalam kondisi pasar yang tidak stabil, keputusan investor ritel sering kali lebih didorong oleh respons emosional, seperti *loss aversion*, yang sulit diprediksi oleh model ekonomi tradisional [2]. Lebih lanjut, data pasar Indonesia, yang berintegrasi dinamis dengan pasar global, menunjukkan bahwa kondisi makroekonomi domestik seperti inflasi dan nilai tukar memiliki pengaruh signifikan terhadap Indeks Harga Saham Gabungan [3]. Sifat non-stasioner dan *non-linear* dari data *time series* harga saham menjadikan tugas prediksi sangat menantang. Dengan demikian, terdapat kebutuhan mendesak untuk mengembangkan alat prediktif yang lebih canggih dan adaptif dalam memodelkan pergerakan harga saham.

Kebutuhan ini muncul karena kompleksitas data pasar yang tidak dapat ditangani secara optimal oleh metode konvensional. Oleh karena itu, pendekatan berbasis teknologi modern menjadi solusi yang relevan untuk meningkatkan akurasi prediksi.

Penelitian sebelumnya dalam prediksi harga saham secara tradisional banyak mengandalkan model ekonometri seperti *Autoregressive Integrated Moving Average* (ARIMA) dan *Generalized Autoregressive Conditional Heteroskedasticity* (GARCH). Namun, model-model ini sering kali menunjukkan keterbatasan fundamental dalam menangkap pola *non-linear* dan kompleksitas inheren dari data pasar modal, sebagaimana dicatat oleh banyak literatur. Kesenjangan ini mendorong transisi ke model berbasis Kecerdasan Buatan. Studi terdahulu telah mengeksplorasi penggunaan Jaringan Saraf Tiruan (ANN) dan *Long Short-Term Memory* (LSTM), bahkan beberapa telah mengintegrasikannya dengan model statistik seperti GARCH [4] menunjukkan superioritas model hibrida dan AI. Namun, tinjauan terhadap model *deep learning* juga mengungkapkan tantangan dalam hal interpretasi dan kebutuhan komputasi yang tinggi [5]. Kesenjangan utamanya terletak pada eksplorasi dan perbandingan kinerja model *ensemble learning* berkinerja tinggi seperti Random Forest dan *Light Gradient Boosting Machine* (LightGBM) secara langsung terhadap data pasar saham spesifik di Indonesia, yang dikenal memiliki karakteristik unik yang belum sepenuhnya teratasi.

Tujuan utama dari penelitian ini adalah untuk membandingkan dan mengevaluasi kinerja dua model *ensemble learning*, yaitu Random Forest dan LightGBM, dalam melakukan prediksi harga saham harian di Bursa Efek Indonesia (BEI). Dengan demikian, penelitian ini bertujuan untuk mengidentifikasi model mana yang memberikan tingkat akurasi dan efisiensi komputasi yang lebih optimal dalam konteks pasar negara berkembang. Signifikansi penelitian ini terletak pada kontribusinya untuk mengisi kesenjangan literatur mengenai penerapan algoritma *ensemble* lanjutan di pasar Asia Tenggara yang memiliki dinamika pasar yang berbeda dari pasar maju. Selain itu, penelitian ini memberikan kontribusi praktis yang penting bagi investor ritel dan institusi di Indonesia

dengan menyediakan basis data yang relevan dan dapat diakses [6], serta model prediksi yang dapat dijadikan panduan dalam pengambilan keputusan investasi yang lebih terinformasi.

Pendekatan yang akan digunakan dalam penelitian ini adalah metode kuantitatif dengan fokus pada implementasi dan evaluasi algoritma Pembelajaran Mesin. *Ensemble learning*, khususnya Random Forest (RF) dan LightGBM, dipilih karena kemampuannya yang telah terbukti dalam menangani data *non-linear* dan data dengan dimensi tinggi, yang merupakan karakteristik khas dari data finansial. LightGBM, sebagai pengembangan dari *Gradient Boosting Decision Tree* (GBDT), terkenal karena efisiensi komputasi yang tinggi dan akurasi yang unggul dibandingkan model lain dalam skenario *time series* [7]. Hasil empiris di pasar yang serupa (misalnya pasar Tiongkok) telah menunjukkan bahwa LightGBM memberikan kemampuan prediksi yang jauh lebih baik daripada regresi tradisional [8]. Lebih lanjut, kombinasi LightGBM dalam model hibrida juga telah terbukti meningkatkan akurasi prakiraan secara signifikan [9]. Oleh karena itu, melalui serangkaian proses *preprocessing* data, pelatihan model, dan optimasi *hyperparameter*, kedua algoritma akan diuji secara ketat untuk menentukan model yang paling *robust* untuk prediksi harga saham di BEI.

Keberhasilan penelitian ini dalam mengidentifikasi model prediksi yang paling akurat akan memiliki implikasi substansial, baik bagi bidang Informatika maupun sektor finansial. Jika LightGBM terbukti superior, hal ini akan memperkuat argumentasi bahwa model berbasis *gradient boosting* merupakan alat yang paling adaptif dan responsif terhadap perubahan kondisi pasar yang kompleks di *emerging market*. Temuan ini akan berkontribusi pada pemahaman yang lebih baik tentang bagaimana kompleksitas *non-linear* data finansial dapat dimodelkan secara efektif menggunakan pendekatan *ensemble learning*. Secara praktis, hasil penelitian ini dapat digunakan sebagai dasar bagi para pengembang sistem informasi dan *fintech* untuk merancang aplikasi analisis pasar yang lebih canggih. Selain itu, model yang dihasilkan dapat menjadi referensi utama bagi para praktisi investasi untuk mengintegrasikan analitik prediktif berbasis AI. Kontribusi utama riset ini adalah menyediakan solusi berbasis *data-driven* yang kredibel, yang pada akhirnya dapat meningkatkan literasi finansial dan kemampuan pengambilan keputusan berbasis bukti di pasar modal Indonesia.

MATERIALS AND METHODS

Alur prosesnya saya rancang menggunakan diagram untuk memvisualisasikan langkah-langkahnya, dan saya jelaskan secara naratif agar lebih mudah dipahami. Mulai dari tahap awal, yaitu mengidentifikasi masalah yang ingin diselesaikan dan menetapkan tujuan penelitian. Setelah itu, saya lanjut ke pengumpulan data, membersihkannya dari noise atau kesalahan, lalu membentuk fitur-fitur yang relevan. Data kemudian dipisah menjadi bagian *training* dan *testing*, sebelum membangun model *machine learning*. Langkah berikutnya adalah mengevaluasi performa model dan menganalisis hasilnya secara mendalam. Diagram alur itu sendiri menunjukkan bagaimana setiap proses saling terhubung, termasuk adanya feedback loop untuk melakukan iterasi pada model dan tuning parameter agar lebih optimal. Saya tekankan pentingnya transparansi dan logika berurutan dalam setiap langkah, sesuai dengan panduan eksperimental dari [10] dan [11]. Saya cantumkan pada Gambar 1.



Gambar 1 Alur Penelitian

Pengumpulan Data Pengambilan data saham yang bersifat *cross-sectional* berbasis karakteristik fundamental perusahaan dari Bursa Efek Indonesia (BEI)

Pembersihan dan Persiapan Data Data mentah dibersihkan dari nilai hilang, outlier, dan inkonsistensi. Transformasi dilakukan untuk normalisasi, *encoding* fitur kategorikal, serta

penyusunan fitur baru berdasarkan domain keuangan.

Pembagian Data Data yang telah diproses kemudian dibagi menjadi set pelatihan, validasi, dan pengujian. Teknik pembagian mengikuti pendekatan berbasis waktu (chronological split) dengan skema rolling window agar menghindari kebocoran data.

Pemodelan Pembangunan model prediksi dilakukan dengan dua algoritma utama: Random Forest dan LightGBM. Model dikembangkan menggunakan parameter default yang kemudian dioptimasi untuk hasil terbaik.

Evaluasi Model Performa model diuji menggunakan metrik evaluasi seperti MAE, RMSE, dan R^2 . Hasil ini membantu mengukur ketepatan dan kemampuan generalisasi model terhadap data baru.

Analisis Feature Importance Setelah model dibangun dan dievaluasi, dilakukan analisis untuk mengetahui fitur mana yang paling berpengaruh terhadap prediksi. Teknik seperti permutation importance dan SHAP digunakan dalam tahap ini.

RESULTS AND DISCUSSION

Pada gambar 1 pemodelan, dilakukan pelatihan model menggunakan dua algoritma, yaitu Random Forest dan Light Gradient Boosting Machine (LightGBM). Proses pelatihan diawali dengan model Random Forest yang berhasil dilatih dengan waktu komputasi sebesar 0,288 detik. Hal ini menunjukkan bahwa model mampu dibangun dengan cukup cepat dan efisien pada dataset yang digunakan. selanjutnya dilakukan pelatihan menggunakan algoritma LightGBM.

Berdasarkan hasil proses pelatihan, model ini memiliki waktu komputasi yang lebih cepat, yaitu sebesar 0,034 detik. Hal ini menunjukkan bahwa LightGBM menunjukkan efisiensi komputasi yang lebih baik, namun belum sepenuhnya mampu menghasilkan model prediksi yang optimal dibandingkan Random Forest, terutama dalam menangani dataset dengan jumlah fitur yang relatif banyak. Untuk proses pelatihan LightGBM, muncul beberapa peringatan (*warning*), seperti “No further splits with positive gain, best gain: -inf”. Peringatan ini menunjukkan bahwa pada beberapa iterasi, model tidak menemukan pembagian data (*split*)

yang dapat meningkatkan nilai informasi (*gain*). Dengan kata lain, model mengalami kesulitan dalam menemukan pola tambahan yang signifikan dari data yang tersedia. Terdapat informasi bahwa jumlah data pelatihan yang digunakan sebanyak 659 dengan jumlah fitur sebanyak 13. Hal ini mengindikasikan bahwa model bekerja pada dataset dengan dimensi yang cukup terbatas, sehingga kemungkinan informasi yang dapat diekstraksi juga terbatas. Peringatan lain yang muncul terkait dengan penamaan fitur (*feature names*) menunjukkan bahwa terdapat penggunaan spasi dalam nama variabel yang kemudian disarankan untuk diganti dengan garis bawah (*underscore*). Meskipun tidak mempengaruhi hasil secara signifikan, hal ini merupakan praktik yang baik untuk menjaga kompatibilitas dan efisiensi proses komputasi.

Secara keseluruhan, proses pemodelan menunjukkan bahwa kedua algoritma berhasil dilatih dengan baik. Namun demikian, terdapat indikasi bahwa model LightGBM belum sepenuhnya optimal dalam menangkap pola yang terdapat dalam data. Hal ini menunjukkan bahwa masih diperlukan penyesuaian lebih lanjut agar performa model dapat ditingkatkan secara maksimal.



Gambar 1 Pemodelan

HASIL EVALUASI MODEL

Setelah proses pemodelan dan penentuan hyperparameter dilakukan, tahap selanjutnya adalah mengevaluasi kinerja model menggunakan data uji. Evaluasi ini bertujuan untuk mengukur sejauh mana model mampu melakukan prediksi secara akurat terhadap data yang belum pernah dilihat sebelumnya. Berdasarkan hasil pengujian, model Random Forest menghasilkan nilai *Root Mean Squared Error* (RMSE) sebesar 2949.26 dan *Mean Absolute Error* (MAE) sebesar 1675.10. Sementara itu, nilai koefisien determinasi (R^2) yang diperoleh adalah -2.1227. Nilai R^2 yang bernilai negatif menunjukkan bahwa model Random Forest belum mampu menangkap pola hubungan antara variabel independen dan variabel dependen dengan baik, bahkan performanya lebih buruk dibandingkan dengan model sederhana yang hanya menggunakan rata-rata sebagai prediksi.

Untuk model *Light Gradient Boosting Machine* (LightGBM) menunjukkan performa yang relatif lebih baik dibandingkan Random Forest. Model ini menghasilkan nilai RMSE sebesar 2293.67 dan MAE

sebesar 1628.84, yang lebih rendah dibandingkan Random Forest. Hal ini menunjukkan bahwa tingkat kesalahan prediksi yang dihasilkan oleh LightGBM lebih kecil. Selain itu, nilai R^2 yang diperoleh sebesar -0.8887, meskipun masih bernilai negatif, namun lebih mendekati nol dibandingkan dengan Random Forest, yang mengindikasikan bahwa model ini memiliki kemampuan yang lebih baik dalam menjelaskan variasi data.

Dari segi waktu komputasi, LightGBM juga menunjukkan keunggulan dengan waktu pelatihan sebesar 0.034 detik, jauh lebih cepat dibandingkan Random Forest yang membutuhkan waktu sebesar 0.288 detik. Hal ini menunjukkan bahwa LightGBM lebih efisien dalam proses pelatihan model, terutama ketika digunakan pada dataset dengan jumlah fitur yang cukup banyak. Secara keseluruhan, berdasarkan perbandingan nilai RMSE, MAE, R^2 , dan waktu pelatihan, dapat disimpulkan bahwa model LightGBM memiliki performa yang lebih baik dibandingkan Random Forest dalam penelitian ini. Meskipun demikian, kedua model masih menunjukkan nilai R^2 yang negatif, yang mengindikasikan bahwa model belum optimal dalam menangkap pola data secara keseluruhan. Hal ini dapat disebabkan oleh beberapa faktor, seperti karakteristik data yang kompleks, kurangnya fitur yang relevan, atau adanya variabel penting yang tidak dimasukkan dalam model, seperti variabel fundamental atau faktor eksternal yang mempengaruhi harga saham. Oleh karena itu, diperlukan pengembangan lebih lanjut, seperti penambahan fitur, penggunaan teknik *feature engineering*, atau pendekatan model yang lebih kompleks untuk meningkatkan kinerja prediksi.

Tabel 1 Hasil Evaluasi RF

	Metric	Value
0	RMSE	2949.26
1	MAE	1675.10
2	R2	-2.1227
3	Waktu (s)	0.288

Tabel 2 Hasil Evaluasi LightGBM

	Metric	Value
0	RMSE	2293.67
1	MAE	1628.84
2	R2	-0.8887
3	Waktu (s)	0.034

Tabel 3 Ringkasan perbandingan

Model	RMSE	MAE	R2	Time (s)
-------	------	-----	----	----------

0	Random Forest	2949.26	1675.10	-2.1227	0.288
1	LightGBM	2293.67	1628.84	-0.8887	0.034

RMSE adalah metrik untuk mengukur seberapa besar rata-rata kesalahan prediksi model terhadap nilai sebenarnya. RMSE memberikan penalti lebih besar pada error yang besar sehingga sensitif terhadap *outlier*. Model dianggap semakin baik jika nilai RMSE semakin kecil.

MAE mengukur rata-rata kesalahan absolut antara nilai prediksi dan nilai aktual tanpa memberikan penalti lebih besar pada error besar. MAE lebih stabil dan mudah diinterpretasikan. Model dikatakan lebih baik apabila nilai MAE semakin kecil.

R^2 menunjukkan seberapa besar variasi data aktual yang dapat dijelaskan oleh model. Nilainya berada antara 0 hingga 1, dan semakin mendekati angka 1 berarti model semakin akurat dalam menjelaskan pola data. Model terbaik adalah yang memiliki nilai R^2 paling tinggi.

CONCLUSION

Berdasarkan hasil analisis terhadap data saham dari Bursa Efek Indonesia (BEI) yang mencakup 829 entri dan 14 atribut fundamental, serta pemodelan menggunakan algoritma Random Forest dan *Light Gradient Boosting Machine* (LightGBM), maka dapat disimpulkan beberapa hal berikut:

Penerapan algoritma *machine learning* seperti Random Forest dan LightGBM dapat digunakan untuk membangun model prediksi harga saham berbasis data fundamental perusahaan. Namun, berdasarkan hasil evaluasi, model yang dihasilkan belum menunjukkan performa yang optimal, yang ditunjukkan oleh nilai *koefisien determinasi* (R^2) yang masih bernilai negatif. Hal ini menunjukkan bahwa model belum mampu menangkap hubungan yang kuat antara variabel independen dan variabel dependen.

berdasarkan perbandingan kinerja model, algoritma LightGBM menunjukkan performa yang relatif lebih baik dibandingkan Random Forest, yang ditunjukkan oleh nilai *Mean Absolute Error* (MAE) dan *Root Mean Squared Error* (RMSE) yang lebih rendah. Hal ini mengindikasikan bahwa LightGBM lebih efektif dalam menangani kompleksitas data dibandingkan Random Forest pada penelitian ini.

Hasil analisis *feature importance* menunjukkan bahwa variabel *Shares*, *Sector*, dan *ListingBoard* merupakan faktor yang paling berpengaruh terhadap prediksi harga saham, yang sejalan dengan teori bahwa karakteristik fundamental perusahaan memiliki peran penting dalam menentukan nilai pasar saham.

Penelitian ini memiliki beberapa keterbatasan yang dapat memengaruhi hasil prediksi. Penggunaan data yang bersifat *cross-sectional* menyebabkan model belum mampu menangkap pola temporal dari pergerakan harga saham. Selain itu, variabel yang digunakan masih terbatas pada faktor fundamental perusahaan dan belum mempertimbangkan faktor eksternal seperti kondisi makroekonomi dan sentimen pasar.

Penelitian ini menunjukkan bahwa pendekatan berbasis *data-driven* memiliki potensi untuk digunakan dalam mendukung pengambilan keputusan investasi. Namun, keterbatasan pada jenis data, variabel yang digunakan, serta metode evaluasi menyebabkan hasil model belum dapat digeneralisasi secara luas.

REFERENCE

- [1] M. Kacperczyk, "Global volatility and firm level capital flows," *J. financ. econ.*, vol. 149, no. 2, pp. 279–300, 2025, doi: 10.1016/j.jfineco.2024.02.010.
- [2] D. van Dolder and J. Vandenbroucke, "Behavioral risk profiling: Measuring loss aversion of individual investors," *J. Bank. Financ.*, vol. 168, p. 107293, 2024, doi: 10.1016/j.jbankfin.2024.107293.
- [3] I. Yuliadi, S. P. Murpratomo, and D. T. K. Wardani, "Integration of Stock Markets Between Indonesia and Its Major Trading Partners," *Journal of Economics Research and Social Sciences*, vol. 8, no. 2, pp. 277–289, 2024, doi: 10.18196/jerss.v8i2.22154.
- [4] J. Li, J. K. Mutinda, and A. K. Langat, "Stock Price Prediction Using Combined GARCH-AI Models," *Sci. Afr.*, vol. 26, 2024, doi: 10.1016/j.sciaf.2024.e02374.
- [5] W. Bao, Y. Cao, Y. Yang, H. Che, J. Huang, and S. Wen, "Data-driven stock forecasting models based on neural networks: A review," *Information Fusion*, vol. 113, p. 102616, 2025, doi: 10.1016/j.inffus.2024.102616.
- [6] J. Hartanto, T. Liundi, R. Sutoyo, and E. W. Andangsari, "ID-SMSA: Indonesian stock market dataset for sentiment analysis," *Data Brief*, vol. 60, p. 111571, 2025, doi: 10.1016/j.dib.2025.111571.
- [7] A. D. Hartanto, Y. N. Kholik, and Y. Pristyanto, "Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model," *International Journal on Informatics Visualization*, vol. 7, no. 4, pp. 2270–2279, 2023.
- [8] J. Yang, "Application of LightGBM in the Chinese stock market," in *2025 4th International Conference on Big Data, Information and Computer Network (BDICN 2025)*, Guangzhou, China, 2025. doi: 10.1145/3727353.3727490.
- [9] A. Ampountolas and S. AlGharbi, "An innovative hybrid LightGBM-BPNN model for enhanced commodity forecasting accuracy," *Finance Research Open*, vol. 1, p. 100004, 2025, doi: 10.1016/j.finr.2025.100004.
- [10] M. Rezaei and P. Johnson, "Best-practice roadmaps for machine learning experimentation BT - Proceedings of the 2025 IEEE International Conference on Data Science," 2025. doi: 10.1109/ICDS.2025.00045.
- [11] H. Gao *et al.*, "Machine learning in business and finance: A literature review and research opportunities," *Inf. Syst.*, vol. 10, p. 86, 2024, doi: 10.1016/j.is.2024.102456.