Evaluasi Stabilitas Performa Model *Machine Learning* dalam Prediksi *Return* IHSG pada Rezim Volatilitas Berbeda



Disusun Oleh:

Nama : R. Aristyo Naufal Triyoga

NIM : 22523275

PROGRAM STUDI INFORMATIKA – PROGRAM SARJANA

FAKULTAS TEKNOLOGI INDUSTRI

UNIVERSITAS ISLAM INDONESIA

2026

HALAMAN PENGESAHAN DOSEN PEMBIMBING

Evaluasi Stabilitas Performa Model *Machine Learning* dalam Prediksi *Return* IHSG pada Rezim Volatilitas Berbeda

TUGAS AKHIR

Disusun Oleh:

Nama : R. Aristyo Naufal Triyoga
NIM : 22523275

Yogyakarta, 1 Juli 2026
Pembimbing,

(Arrie Kurniawardhani, S.Si., M.Kom.)

HALAMAN PENGESAHAN DOSEN PENGUJI

Evaluasi Stabilitas Performa Model *Machine Learning* dalam Prediksi *Return* IHSG pada Rezim Volatilitas Berbeda

TUGAS AKHIR

Telah dipertahankan di depan sidang penguji sebagai salah satu syarat untuk memperoleh gelar Sarjana Komputer dari Program Studi Informatika – Program Sarjana di Fakultas Teknologi Industri Universitas Islam Indonesia

Yogyakarta, 27 Juli 2026

Tim Penguji

Arrie Kurniawardhani, S.Si., M.Kom. _____

Anggota 1

Feri Wijayanto, Dr., S.T., M.T. _____

Anggota 2

Moh. Idris, S.Kom., M.Kom. _____

Mengetahui,
Ketua Program Studi Informatika – Program Sarjana
Fakultas Teknologi Industri
Universitas Islam Indonesia

(Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D.)

HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan di bawah ini:

Nama : R. Aristyo Naufal Triyoga

NIM : 22523275

Tugas akhir dengan judul:

Evaluasi Stabilitas Performa Model *Machine Learning* dalam Prediksi *Return* IHSG pada Rezim Volatilitas Berbeda

Menyatakan bahwa seluruh komponen dan isi dalam tugas akhir ini adalah hasil karya saya sendiri. Apabila di kemudian hari terbukti ada beberapa bagian dari karya ini adalah bukan hasil karya sendiri, tugas akhir yang diajukan sebagai hasil karya sendiri ini siap ditarik kembali dan siap menanggung risiko dan konsekuensi apapun.

Demikian surat pernyataan ini dibuat, semoga dapat dipergunakan sebagaimana mestinya.

Yogyakarta, 1 Juli 2026



(R. Aristyo Naufal Triyoga)

HALAMAN PERSEMBAHAN

Alhamdulillahirabbil'alamin. Puji syukur tak terhingga penulis panjatkan ke hadirat Allah *Subhanahu Wa Ta'ala*. Atas kehendak, rahmat, dan kelapangan yang dianugerahkan-Nya, proses panjang dalam penyusunan tugas akhir ini akhirnya dapat diselesaikan.

Penulis menyadari bahwa perjalanan ini tidak mungkin sampai pada titik akhir tanpa doa, cinta, dukungan, dan kehadiran orang-orang yang sangat berarti. Oleh karena itu, dengan kerendahan hati dan rasa syukur, karya ini penulis persembahkan kepada:

1. Bapak dan Ibu penulis, Aris Wimbargo dan Siti Maryam Valentina, serta kedua kakak penulis, Aristianto Adi Nugroho dan Aristiawan Hanif, atas doa, kasih sayang, kesabaran, dukungan, pengorbanan, serta kepercayaan yang tidak pernah berhenti diberikan kepada penulis. Segala bentuk dukungan tersebut menjadi kekuatan terbesar bagi penulis untuk terus melangkah dan menyelesaikan tugas akhir ini.
2. Paman dan bibi penulis, Susilo Trihardjono dan Siti Katarina, serta sepupu penulis, Karina Dhesy, yang telah memberikan dukungan baik secara materiel maupun imateriel selama proses penyusunan tugas akhir ini. Bantuan, perhatian, dan kepedulian yang diberikan menjadi bagian penting dalam perjalanan penulis hingga dapat menyelesaikan tugas akhir ini.
3. Dendy, Dudut, Gatot, Rangga, Fauzan, Gana, Ariel, Haris, Farid, Hasan, Novera, Widya, Sekar, Adel, serta seluruh teman-teman Invacy yang telah menjadi teman seperjuangan sejak awal masa perkuliahan. Terima kasih atas kebersamaan, bantuan, dukungan, diskusi, serta pengalaman yang telah dibagikan selama perjalanan akademik ini. Seluruh kenangan tersebut menjadi bagian berharga dalam perjalanan hidup penulis.
4. Elang, Raken, Pandu, Fafan, dan Alfat, sebagai teman diskusi, teman berbagi cerita, serta tempat bertukar pikiran di luar lingkungan perkuliahan. Terima kasih atas waktu, dukungan, candaan, dan obrolan yang membantu penulis menjaga semangat selama proses penyusunan tugas akhir ini. Kehadiran dan kebersamaan yang diberikan menjadi ruang bagi penulis untuk beristirahat sejenak dari tekanan akademik, sekaligus kembali menemukan semangat untuk menyelesaikan tanggung jawab ini.

HALAMAN MOTO

"And, when you want something, all the universe conspires in helping you to achieve it."
(Paulo Coelho, *The Alchemist*)

KATA PENGANTAR

Alhamdulillahi Rabbil Alamin. Segala puji bagi Allah *Subhanahu Wa Taala* atas rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan tugas akhir ini. Penulis menyadari bahwa proses penyusunan tugas akhir ini tidak lepas dari berbagai rintangan dan hambatan. Namun, berkat doa, dukungan, arahan, serta bantuan dari berbagai pihak, tugas akhir ini dapat diselesaikan dengan baik. Dengan penuh rasa hormat dan rendah hati, penulis ingin menyampaikan terima kasih kepada:

1. Allah *Subhanahu Wa Ta`ala*, atas segala rahmat, kekuatan, dan kemudahan yang diberikan dalam setiap proses penyusunan tugas akhir ini.
2. Bapak Prof. Dr. Ir. Hari Purnomo, M.T., IPU., ASEAN Eng., selaku Rektor Universitas Islam Indonesia.
3. Ibu Dr. Sri Kusumadewi, S.Si., M.T., selaku Dekan Fakultas Teknologi Industri Universitas Islam Indonesia.
4. Bapak Dr. Raden Teduh Dirgahayu, S.T., M.Sc., selaku Ketua Jurusan Informatika Universitas Islam Indonesia.
5. Bapak Dhomas Hatta Fudholi, S.T., M.Eng., Ph.D., selaku Ketua Program Studi Informatika Universitas Islam Indonesia.
6. Ibu Arrie Kurniawardhani, S.Si., M.Kom., selaku Dosen Pembimbing Skripsi yang telah sabar membimbing, mengarahkan, serta memberikan banyak masukan berharga dalam penyusunan tugas akhir ini.
7. Bapak Zainudin Zukhri, S.T., M.I.T., selaku Dosen Pembimbing Akademik yang telah memberikan arahan dan perhatian selama masa perkuliahan.
8. Seluruh staf dan karyawan Fakultas Teknologi Industri Universitas Islam Indonesia yang telah membantu dalam berbagai proses administrasi dan akademik selama masa perkuliahan.

Penulis juga mengucapkan terima kasih kepada keluarga, sahabat, teman-teman, serta semua pihak yang telah menjadi perantara kebaikan dari Allah *Subhanahu Wa Ta`ala* dalam penyelesaian tugas akhir ini.

Yogyakarta, 1 Juli 2026



(R. Aristyo Naufal Triyoga)

SARI

Penelitian ini mengevaluasi stabilitas performa model *machine learning* dalam memprediksi *return* Indeks Harga Saham Gabungan (IHSG) 20 sesi perdagangan ke depan. Model yang diuji adalah Ridge Regression, Random Forest, LightGBM, dan XGBoost. Penelitian ini menggunakan tiga rancangan dataset, yaitu Dataset A yang berisi fitur teknikal, Dataset B yang menambahkan variabel makroekonomi domestik, dan Dataset C yang menambahkan variabel makro inti serta fitur volatilitas berbasis GARCH. Evaluasi dilakukan menggunakan RMSE dan MAE melalui *walk-forward validation* dengan enam *fold* tahunan 2020 sampai dengan 2025 dan 72 observasi bulanan pada tanggal pengujian yang identik untuk seluruh dataset. Model dibandingkan dengan dua metode pembanding sederhana, yaitu *no-change forecast* dan *persistence forecast*. Perbedaan akurasi diuji menggunakan uji Diebold-Mariano dengan koreksi Holm, sedangkan rezim volatilitas dibentuk menggunakan Markov Switching.

Random Forest menghasilkan Mean RMSE terendah pada ketiga dataset, dengan nilai terendah 0,044536 pada Dataset B. Namun tidak satu pun dari 120 perbandingan antarmodel bertahan setelah koreksi Holm, dengan nilai p terkoreksi terkecil 0,0767, sehingga keunggulan tersebut hanya bersifat deskriptif. Terhadap *persistence forecast*, seluruh dua belas kombinasi menghasilkan RMSE lebih rendah dengan perbaikan antara 8,706% dan 19,392%. Terhadap *no-change forecast*, hanya tiga kombinasi yang lebih rendah dengan perbaikan terbesar 0,996%. Galat prediksi berubah tajam antarperiode uji, yaitu dari rata-rata RMSE 0,109953 pada 2020 menjadi 0,020044 pada 2022. Galat juga tidak meningkat berurutan dari rezim volatilitas rendah ke rezim tinggi karena galat terbesar justru terjadi pada rezim sedang pada sebelas dari dua belas kombinasi. Penelitian ini karena itu tidak menemukan model yang stabil unggul, dan variasi galat yang berasal dari periode pengujian jauh lebih besar daripada variasi yang berasal dari pilihan model.

Kata kunci: IHSG, *return*, *machine learning*, *walk-forward validation*, rezim volatilitas, stabilitas performa.

GLOSARIUM

Ablation Study	Prosedur menilai kontribusi suatu kelompok fitur dengan membandingkan performa model sebelum dan sesudah kelompok fitur tersebut ditambahkan.
Alpha	Parameter pada Ridge Regression yang mengatur kuat atau lemahnya regularisasi.
Chronological Holdout	Pembagian data pelatihan dan data pengujian berdasarkan urutan waktu, dengan data terbaru dipakai untuk menguji.
Coefficient of Variation	Rasio standar deviasi terhadap rata-rata, dipakai untuk membandingkan kestabilan galat antarmodel.
Dataset	Kumpulan data yang digunakan dalam penelitian.
Directional Accuracy	Persentase prediksi yang arah naik atau turunnya sama dengan arah return aktual.
Fold	Satu putaran pelatihan dan pengujian dalam *walk-forward validation*, yang pada penelitian ini mewakili satu tahun pengujian.
GARCH	Metode untuk mengestimasi volatilitas yang berubah dari waktu ke waktu pada data keuangan.
HAC	*Heteroskedasticity and Autocorrelation Consistent*, koreksi ragam agar uji tetap sahih ketika galat prediksi berkorelasi antarwaktu.
Koreksi Holm	Penyesuaian nilai p ketika banyak perbandingan diuji sekaligus, agar peluang menyimpulkan perbedaan yang sebenarnya tidak ada tidak menumpuk.
Machine Learning	Pendekatan komputasi yang memungkinkan model mempelajari pola dari data untuk menghasilkan prediksi.
MAE	*Mean Absolute Error*, rata-rata nilai mutlak selisih antara prediksi dan nilai aktual.
Markov Switching	Model yang menaksir keadaan pasar yang tidak teramati langsung dari data, dipakai untuk membentuk rezim volatilitas.
No-change Forecast	Metode pembanding sederhana yang memprediksi return ke depan sebesar 0%.

Persistence Forecast	Metode pembanding sederhana yang menggunakan return historis terbaru sebagai prediksi return periode berikutnya.
Prapemrosesan	Tahap penyiapan dan pengolahan data sebelum digunakan untuk melatih model.
Return	Persentase perubahan nilai indeks atau aset dalam periode tertentu.
Rezim Volatilitas	Pengelompokan periode pengujian menjadi volatilitas rendah, sedang, dan tinggi.
RMSE	*Root Mean Squared Error*, akar dari rata-rata kuadrat selisih antara prediksi dan nilai aktual.
Time Series	Data yang tersusun berdasarkan urutan waktu.
Training	Proses pelatihan model menggunakan data pelatihan.
Uji Diebold-Mariano	Uji statistik untuk membandingkan akurasi dua prediksi pada data dan periode yang sama.
Volatilitas	Ukuran besar kecilnya perubahan return dalam suatu periode.
Walk-forward Validation	Metode evaluasi data *time series* yang melatih dan menguji model secara berulang mengikuti urutan waktu.
Winsorization	Teknik untuk membatasi nilai ekstrem agar tidak terlalu mendominasi proses pelatihan model.
Worst-fold RMSE	Nilai RMSE terbesar yang dihasilkan model di antara seluruh fold pengujian.

DAFTAR ISI

HALAMAN JUDUL i
HALAMAN PENGESAHAN DOSEN PEMBIMBING ii
HALAMAN PENGESAHAN DOSEN PENGUJI iii
HALAMAN PERNYATAAN KEASLIAN TUGAS AKHIR iv
HALAMAN PERSEMBAHAN v
HALAMAN MOTO vi
KATA PENGANTAR vii
SARI viii
GLOSARIUM ix
DAFTAR ISI xi
DAFTAR TABEL xiii
DAFTAR GAMBAR xiv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 4
1.4 Tujuan Penelitian 5
1.5 Manfaat Penelitian 5
1.6 Sistematika Penulisan 6
BAB II KAJIAN PUSTAKA 7
2.1 Indeks Harga Saham Gabungan dan *Return* Saham 7
2.1.1 Indeks Harga Saham Gabungan 7
2.1.2 *Return* Saham 7
2.2 Tantangan Prediksi Return Saham 8
2.2.1 *Noise* dan Prediktabilitas Rendah 8
2.2.2 Perubahan Kondisi Pasar 9
2.2.3 Model Markov Switching 10
2.3 *Machine learning* untuk Prediksi *Return* Saham 11
2.3.1 Ridge Regression 11
2.3.2 Random Forest 12
2.3.3 LightGBM 13
2.3.4 XGBoost 14
2.4 Fitur Prediksi yang Digunakan 15
2.4.1 Fitur Teknikal 15
2.4.2 Variabel Makroekonomi 16
2.4.3 Fitur Volatilitas Berbasis GARCH 17
2.4.4 Ablation Study 18
2.5 Metode Pembanding Sederhana 19
2.5.1 No-change Forecast 19
2.5.2 Persistence Forecast 20
2.6 Validasi Model pada Data *Time Series* 21
2.6.1 Chronological Holdout 21
2.6.2 Walk-forward Validation 22
2.6.3 Data Leakage dan Look-ahead Bias 22
2.7 Metrik Evaluasi Model 24
2.7.1 RMSE 24
2.7.2 MAE 24

2.7.3 Alasan Tidak Menggunakan MAPE dan R-squared....25
2.8 Penelitian Terdahulu....27
2.9 Posisi Penelitian....31
BAB III METODOLOGI PENELITIAN....34
3.1 Desain Penelitian....34
3.2 Sumber dan Pengambilan Data....36
3.3 Pembentukan Dataset, Fitur, dan Target....38
3.4 Pembagian Data dan Prapemrosesan....40
3.5 Model dan Metode Pembanding....43
3.6 Metrik Evaluasi....46
3.7 Analisis Stabilitas Performa....48
3.8 Uji Diebold-Mariano....52
3.9 *Ablation Study*....53
BAB IV HASIL DAN PEMBAHASAN....55
4.1 Skenario Eksperimen....55
4.2 Ringkasan Performa Model pada Setiap Dataset....56
4.3 Perbandingan dengan Metode Pembanding Sederhana....59
4.4 Evaluasi Tahunan *Walk-Forward Validation*....62
4.5 Evaluasi Berdasarkan Rezim Volatilitas....63
4.6 Uji Diebold-Mariano....65
4.7 Ablation Study....67
4.8 Pembahasan Stabilitas Performa....68
4.9 Ringkasan Temuan Bab....71
BAB V KESIMPULAN DAN SARAN....73
5.1 Kesimpulan....73
5.2 Saran....74
DAFTAR PUSTAKA....77
LAMPIRAN....81

DAFTAR TABEL

Tabel 2.1 Penelitian Terdahulu 27
Tabel 3.1 Ringkasan Sumber Data 37
Tabel 3.2 Kelompok Fitur Teknikal 38
Tabel 3.3 Ringkasan Dataset Penelitian 40
Tabel 3.4 Ringkasan Protokol Evaluasi Utama 40
Tabel 3.5 Pembagian Data Expanding Walk-Forward Validation 41
Tabel 3.6 Konfigurasi dan Penalaan Model 45
Tabel 3.7 Konfigurasi Markov Switching 50
Tabel 3.8 Ringkasan Indikator Stabilitas Performa Model 51
Tabel 3.9 Konfigurasi Uji Diebold-Mariano 53
Tabel 3.10 Rancangan Ablation Study 54
Tabel 4.1 Performa Model 57
Tabel 4.2 Uji Diebold-Mariano Random Forest Dataset B pada Lag HAC 1 66
Tabel 4.3 Sensitivitas Hasil Signifikan terhadap Panjang Lag HAC 66
Tabel 4.4 Perubahan Mean RMSE dan Nilai p Holm Setiap Tahap Ablation 67
Tabel 4.5 Kombinasi Terbaik menurut Kriteria Evaluasi 69

DAFTAR GAMBAR

Gambar 3.1 Alur Eksperimental Penelitian 36
Gambar 3.2 Alur Pembagian Data dan Prapemrosesan pada Setiap Fold 43
Gambar 4.1 Volatilitas Bersyarat GARCH(1,1) Periode 2013 sampai 2026 56
Gambar 4.2 Mean RMSE terhadap Standar Deviasi RMSE Antar-Fold 58
Gambar 4.3 Perbaikan RMSE Setiap Model terhadap No-change Forecast 60
Gambar 4.4 Return Aktual dan Prediksi Random Forest Dataset B 61
Gambar 4.5 RMSE Tahunan Setiap Model pada Enam Fold Pengujian 62
Gambar 4.6 RMSE Setiap Model pada Tiga Rezim Volatilitas 64

BAB I
PENDAHULUAN

1.1 Latar Belakang

Indeks Harga Saham Gabungan (IHSG) merupakan salah satu indikator yang digunakan untuk membaca pergerakan pasar saham Indonesia. Perubahan nilai IHSG dapat menunjukkan kondisi umum pasar, sehingga sering menjadi perhatian investor, analis, dan pelaku pasar. Dalam konteks tersebut, prediksi terhadap perubahan IHSG dapat menjadi salah satu bahan analisis untuk memahami kemungkinan arah perubahan pasar. Namun, hasil prediksi tetap perlu dipahami sebagai estimasi yang memiliki galat, bukan sebagai kepastian atau rekomendasi transaksi (Ismanto, 2026).

Penelitian ini berfokus pada prediksi *return* IHSG. *Return* menunjukkan persentase perubahan nilai indeks dalam periode tertentu. *Return* digunakan karena dapat menggambarkan perubahan relatif, bukan hanya perubahan angka indeks secara langsung. Sebagai contoh, kenaikan IHSG dari 7.000 ke 7.350 dapat dibaca sebagai kenaikan sekitar 5%. Dengan menggunakan *return*, hasil prediksi dapat dibaca sebagai besar perubahan relatif terhadap nilai awal (Indra et al., 2025).

Prediksi *return* saham bukan tugas yang mudah. Data pasar keuangan sering mengandung *noise*, yaitu perubahan yang terlihat acak dan tidak selalu membentuk pola yang jelas. Selain itu, kondisi pasar dapat berubah dari waktu ke waktu, sehingga model yang bekerja baik pada satu periode belum tentu menghasilkan performa yang sama pada periode lain. Oleh karena itu, model prediksi perlu diuji secara hati-hati menggunakan data yang belum pernah digunakan dalam proses pelatihan (Fama, 1995; Indra et al., 2025)

Penelitian prediksi IHSG sudah cukup banyak dilakukan, tetapi sebagian besar menargetkan tingkat indeks, bukan *return*. Prediksi IHSG menggunakan regresi k-Nearest Neighbors pada data harian tahun 2022 melaporkan *Mean Absolute Percentage Error* di bawah 2% (Najibullah et al., 2024). Prediksi pergerakan IHSG menggunakan Support Vector Regression pada data mingguan April 2022 sampai Oktober 2024 *melaporkan Mean Absolute Percentage Error* sebesar 1,43% dengan RMSE 121,6196 dalam satuan poin indeks (Galena et al, 2025) Galat persentase yang kecil pada kedua penelitian tersebut berkaitan dengan target berupa tingkat indeks yang bergerak pada kisaran ribuan poin, sehingga angkanya tidak dapat dibandingkan langsung dengan galat prediksi *return* yang bergerak pada kisaran persen.

Perbedaan target tersebut menjadi salah satu alasan penelitian ini memilih *return* sebagai target prediksi.

Machine learning dapat digunakan untuk membangun model prediksi karena model belajar dari pola pada data historis. Dalam konteks pasar saham, pendekatan ini perlu diuji secara hati-hati karena pola historis tidak selalu berulang pada data baru. Model yang lebih kompleks juga belum tentu menghasilkan galat lebih rendah, terutama ketika sinyal prediktif pada data *return* lemah. Karena itu, penelitian ini membandingkan beberapa model dengan karakteristik berbeda, yaitu Ridge Regression, Random Forest, LightGBM, dan XGBoost (Cao, 2025; Hewamalage et al., 2023).

Informasi di luar data harga juga berpotensi berkaitan dengan pergerakan IHSG. Penelitian pada data bulanan Juni 2007 sampai Desember 2021 menemukan bahwa suku bunga, inflasi, nilai tukar, dan harga minyak dunia berkaitan dengan IHSG, tetapi arah dan kekuatan hubungan tersebut berbeda antara periode sebelum dan selama pandemi COVID-19 (Marpaung & Pangestuti, 2024). Harga minyak dunia, misalnya, berpengaruh signifikan terhadap IHSG sebelum pandemi tetapi tidak lagi berpengaruh setelahnya. Temuan tersebut merupakan analisis hubungan antarvariabel, bukan pengujian akurasi prediksi, sehingga tidak dapat langsung dibaca sebagai bukti bahwa variabel makroekonomi meningkatkan akurasi prediksi *return*. Meskipun demikian, temuan itu menjadi dasar penelitian ini untuk menguji apakah penambahan indikator makroekonomi berpengaruh terhadap galat prediksi.

Selain informasi makroekonomi, karakteristik volatilitas IHSG juga menjadi perhatian penelitian terdahulu. Data IHSG menunjukkan *volatility clustering*, yaitu periode bergejolak cenderung diikuti periode bergejolak, sehingga penelitian terdahulu memakai model yang mengizinkan ragam berubah sepanjang waktu seperti ARCH dan GARCH (Adinda et al., 2026). Pendekatan hibrida yang menggabungkan model ragam bersyarat dengan jaringan saraf juga sudah diterapkan pada data IHSG untuk memprediksi volatilitasnya (Wiratama et al., 2025). Kedua penelitian tersebut menargetkan harga dan volatilitas, bukan *return* ke depan. Penelitian ini menggunakan informasi volatilitas dengan cara yang berbeda, yaitu sebagai fitur tambahan pada salah satu rancangan dataset dan sebagai dasar pengelompokan kondisi pasar pada tahap evaluasi, bukan sebagai target prediksi.

Selain membandingkan model *machine learning* satu dengan yang lain, penelitian ini juga menggunakan metode pembanding sederhana. Metode pembanding sederhana diperlukan untuk melihat apakah model *machine learning* memberikan nilai tambah dibandingkan aturan

prediksi dasar. Metode pembanding yang digunakan adalah *no-change forecast* dan *persistence forecast (Beck et al., 2025; Campbell & Thompson, 2008)*.

Evaluasi model pada data pasar saham perlu mempertahankan urutan waktu karena prediksi selalu dibuat berdasarkan informasi masa lalu. Oleh karena itu, penelitian ini menggunakan evaluasi berbasis waktu agar hasil pengujian tidak bergantung pada pembagian data acak (Deep et al., 2025; Hewamalage et al., 2023).

Performa model juga dapat berbeda pada kondisi pasar yang lebih tenang atau lebih bergejolak. Oleh karena itu, penelitian ini menggunakan rezim volatilitas rendah, sedang, dan tinggi sebagai kelompok evaluasi, bukan sebagai target prediksi (G. Deep et al., 2025).

Tiga hal masih jarang dilakukan pada penelitian prediksi *return* saham. Pertama, perbedaan akurasi antarmodel sering dilaporkan hanya sebagai selisih nilai galat tanpa pengujian statistik, padahal selisih yang kecil dapat berasal dari variasi acak. Kedua, kontribusi setiap kelompok fitur jarang diukur secara terpisah, sehingga sulit diketahui bagian mana yang benar-benar berpengaruh. Ketiga, dataset dengan susunan fitur berbeda kadang dievaluasi pada periode yang tidak sama, sehingga perbedaan hasil dapat berasal dari perbedaan periode, bukan dari perbedaan fitur. Penelitian yang menggabungkan validasi berjalan, pengujian statistik dengan koreksi perbandingan ganda, dan *ablation study* menunjukkan bahwa bukti performa yang semula tampak meyakinkan dapat gugur setelah koreksi diterapkan (Pagliaro, 2026). Penelitian berbasis rezim pasar juga menemukan bahwa memasukkan informasi rezim ke dalam model tidak selalu memperbaiki hasil (Gomez Grau, 2025). Ketiga hal tersebut menjadi celah yang ingin dijawab penelitian ini.

Berdasarkan permasalahan tersebut, penelitian ini mengevaluasi performa beberapa model *machine learning* dalam memprediksi *return* IHSG 20 sesi perdagangan ke depan. Penelitian ini menggunakan tiga rancangan dataset untuk membandingkan pengaruh fitur teknikal, informasi makroekonomi, dan informasi volatilitas. Evaluasi dilakukan menggunakan RMSE dan MAE, kemudian dianalisis berdasarkan periode waktu dan rezim volatilitas. Dengan demikian, penelitian ini tidak hanya mencari model dengan galat rata-rata paling rendah, tetapi juga melihat stabilitas performa model pada beberapa kondisi pengujian.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah dalam penelitian ini adalah sebagai berikut.

a. Model *machine learning* mana yang menghasilkan galat prediksi paling rendah dalam memprediksi *return* IHSG 20 sesi perdagangan ke depan?
b. Bagaimana pengaruh penggunaan Dataset A, Dataset B, dan Dataset C terhadap galat prediksi model?
c. Apakah model *machine learning* menghasilkan galat prediksi yang lebih rendah dibandingkan metode pembanding sederhana, yaitu *no-change forecast* dan *persistence forecast?*
d. Bagaimana perubahan galat prediksi model pada setiap periode uji tahunan dalam *walk-forward validation*?
e. Bagaimana perubahan galat prediksi model pada rezim volatilitas rendah, sedang, dan tinggi?

1.3 Batasan Masalah

Agar penelitian tetap terarah, batasan masalah dalam penelitian ini adalah sebagai berikut.

a. Objek penelitian ini adalah Indeks Harga Saham Gabungan (IHSG). Penelitian tidak membahas saham individual, kelompok saham tertentu, atau pembentukan portofolio saham.
b. Target prediksi yang digunakan adalah *return* IHSG 20 sesi perdagangan ke depan. Penelitian ini tidak memprediksi harga IHSG secara langsung dan tidak menguji target *return* harian, mingguan, atau horizon prediksi lain.
c. Dataset yang digunakan dibatasi pada tiga rancangan dataset, yaitu Dataset A, Dataset B, dan Dataset C. Dataset A berisi fitur teknikal IHSG, Dataset B menambahkan indikator makroekonomi domestik, sedangkan Dataset C menambahkan indikator pasar global dan fitur volatilitas berbasis GARCH.
d. Model yang diuji dibatasi pada Ridge Regression, Random Forest, LightGBM, dan XGBoost. Pemilihan model tersebut digunakan untuk membandingkan model linear regularisasi, model *tree ensemble*, dan model *boosting*.
e. Metode pembanding sederhana yang digunakan adalah *no-change forecast* dan *persistence forecast*. Kedua metode tersebut digunakan untuk menilai apakah model *machine learning* memberikan galat prediksi yang lebih rendah dibandingkan aturan prediksi sederhana.
f. Evaluasi model dilakukan menggunakan *Root Mean Squared Error* (RMSE) sebagai metrik utama, dengan *Mean Absolute Error* (MAE) dan *Directional Accuracy* sebagai

metrik pendukung. MAPE dan R-squared tidak digunakan karena kurang sesuai dengan fokus evaluasi pada galat prediksi *return*.

g. Validasi model dilakukan dengan mempertahankan urutan waktu data melalui *expanding walk-forward validation*. Penelitian ini tidak menggunakan pembagian data acak karena data IHSG merupakan data runtun waktu.
h. Perbedaan akurasi antarmodel dan antardataset diuji menggunakan uji Diebold-Mariano dengan koreksi Holm, sedangkan kontribusi setiap kelompok fitur dinilai melalui *ablation study*. Kedua analisis tersebut digunakan untuk menilai kekuatan bukti dan kontribusi fitur pada sampel evaluasi ini, bukan untuk menguji hubungan sebab-akibat.
i. Rezim volatilitas rendah, sedang, dan tinggi digunakan untuk mengevaluasi perubahan galat model pada kondisi pasar yang berbeda. Rezim volatilitas tidak digunakan sebagai target prediksi atau tugas klasifikasi.
j. Penelitian ini berfokus pada evaluasi galat prediksi *return*. Hasil penelitian tidak digunakan untuk menyimpulkan profitabilitas investasi, menyusun strategi *trading*, atau memberikan rekomendasi jual beli.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dijelaskan, tujuan penelitian ini adalah sebagai berikut.

a. Membandingkan performa Ridge Regression, Random Forest, LightGBM, dan XGBoost dalam memprediksi *return* IHSG 20 sesi perdagangan ke depan.
b. Menganalisis pengaruh penggunaan Dataset A, Dataset B, dan Dataset C terhadap galat prediksi model.
c. Membandingkan performa model *machine learning* dengan metode pembanding sederhana, yaitu *no-change forecast* dan *persistence forecast*.
d. Menganalisis perubahan galat prediksi model pada setiap periode uji tahunan dalam *walk-forward validation*.
e. Menganalisis perubahan galat prediksi model pada rezim volatilitas rendah, sedang, dan tinggi.

1.5 Manfaat Penelitian

a. Manfaat Teoritis, penelitian ini diharapkan dapat menambah pembahasan mengenai evaluasi model *machine learning* untuk prediksi *return* IHSG. Pembahasan tersebut

mencakup perbandingan beberapa model, perbandingan beberapa rancangan dataset, serta analisis perubahan galat model pada periode waktu dan rezim volatilitas yang berbeda.

b. Manfaat Praktis, penelitian ini diharapkan dapat membantu pembaca memahami bahwa hasil prediksi model perlu dibandingkan dengan metode pembanding sederhana. Selain itu, penelitian ini juga menunjukkan bahwa galat prediksi dapat berubah ketika kondisi pasar berbeda. Hasil penelitian ini tidak ditujukan sebagai rekomendasi beli, jual, atau menahan saham.
c. Manfaat Metodologis, penelitian ini diharapkan dapat memberikan contoh rancangan evaluasi model prediksi *return* berbasis data runtun waktu. Evaluasi dilakukan dengan mempertahankan urutan waktu data, membandingkan model dengan *no-change forecast* dan *persistence forecast*, serta menganalisis galat berdasarkan rezim volatilitas rendah, sedang, dan tinggi.

1.6 Sistematika Penulisan

Struktur laporan penelitian ini disusun sebagai berikut.

a. Bab I Pendahuluan, menyajikan latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
b. Bab II Kajian Pustaka, menyajikan teori dan penelitian terdahulu yang berkaitan dengan IHSG dan *return* saham, tantangan prediksi *return* saham termasuk rezim volatilitas, model *machine learning*, fitur prediksi yang digunakan beserta *ablation study*, metode pembanding sederhana, validasi model pada data *time series*, metrik evaluasi beserta uji signifikansi perbedaan akurasi, penelitian terdahulu, serta posisi penelitian.
c. Bab III Metodologi Penelitian, menyajikan tahapan penelitian yang meliputi desain penelitian, sumber data, pembentukan dataset, penentuan target prediksi, pembagian data, prapemrosesan, model yang digunakan, metode pembanding sederhana, metrik evaluasi, analisis stabilitas performa model, uji Diebold-Mariano, serta *ablation study*.
d. Bab IV Hasil dan Pembahasan, menyajikan hasil eksperimen dan pembahasan performa model. Bagian ini mencakup perbandingan model pada setiap dataset, perbandingan dengan metode pembanding sederhana, evaluasi tahunan, evaluasi berdasarkan rezim volatilitas, uji Diebold-Mariano, *ablation study*, serta pembahasan stabilitas performa model.
e. Bab V Kesimpulan dan Saran, menyajikan kesimpulan dari hasil penelitian serta saran untuk penelitian selanjutnya.

BAB II
KAJIAN PUSTAKA

2.1 Indeks Harga Saham Gabungan dan *Return* Saham

2.1.1 Indeks Harga Saham Gabungan

Indeks Harga Saham Gabungan (IHSG) merupakan indeks yang digunakan untuk menggambarkan pergerakan pasar saham Indonesia secara umum. IHSG sering dipakai sebagai tolok ukur untuk melihat kinerja pasar saham secara keseluruhan karena nilainya mencerminkan pergerakan saham yang tercatat di Bursa Efek Indonesia. Dengan demikian, IHSG dapat digunakan sebagai objek penelitian ketika penelitian tidak berfokus pada satu saham perusahaan tertentu, tetapi pada gambaran umum pasar saham Indonesia (Ismanto, 2026).

Pergerakan IHSG tidak selalu stabil dari waktu ke waktu. IHSG dapat mengalami perubahan tren, fluktuasi jangka pendek, dan kondisi yang lebih bergejolak. Kondisi tersebut membuat pergerakan IHSG tidak mudah diprediksi secara sederhana. Oleh karena itu, penelitian yang menggunakan IHSG sebagai objek perlu mempertimbangkan bahwa data indeks saham memiliki sifat dinamis dan perlu dievaluasi dengan metode yang hati-hati (Ismanto, 2026).

Dalam penelitian ini, IHSG dipilih sebagai objek karena penelitian berfokus pada indeks pasar, bukan pada saham perusahaan tertentu. Dengan menggunakan IHSG, hasil evaluasi diarahkan untuk membaca performa prediksi pada data indeks pasar secara umum.

2.1.2 *Return* Saham

Return saham menunjukkan perubahan nilai saham atau indeks dalam bentuk persentase. Dalam penelitian ini, *return* digunakan untuk membaca perubahan nilai IHSG secara relatif, bukan hanya perubahan angka indeks secara langsung. Sebagai contoh, kenaikan IHSG dari 7.000 menjadi 7.350 dapat dibaca sebagai kenaikan sebesar 5%. Dengan menggunakan *return*, perubahan IHSG dapat dibandingkan dalam bentuk persentase sehingga lebih mudah dibaca sebagai besar perubahan relatif terhadap nilai awal (Indra et al., 2025).

Penelitian ini tidak memprediksi harga IHSG secara langsung. Target yang diprediksi adalah *return* IHSG untuk 20 sesi perdagangan ke depan. Pemilihan target *return* sejalan dengan penelitian terdahulu pada saham Indonesia yang menggunakan *return* sebagai target prediksi dalam beberapa horizon waktu, termasuk 5 hari dan 21 hari. Horizon 21 hari dalam

penelitian tersebut digunakan untuk merepresentasikan periode bulanan, sedangkan penelitian ini menggunakan 20 sesi perdagangan sebagai pendekatan sekitar satu bulan perdagangan bursa (Beck et al., 2025; Indra et al., 2025).

Dalam penelitian ini, *return* IHSG 20 sesi perdagangan ke depan dihitung dari perbandingan harga penutupan IHSG pada sesi saat prediksi dibuat dengan harga penutupan IHSG 20 sesi perdagangan setelahnya. Rumus perhitungan target prediksi dijelaskan pada Bab 3 karena bagian tersebut termasuk dalam rancangan metodologi penelitian.

Target tersebut berbentuk nilai kontinu. Artinya, model menghasilkan nilai prediksi *return*, bukan memilih kelas seperti naik atau turun. Oleh karena itu, penelitian ini termasuk dalam masalah regresi dan hasil prediksi dievaluasi berdasarkan besar galat prediksi.

2.2 Tantangan Prediksi Return Saham

2.2.1 *Noise* dan Prediktabilitas Rendah

Prediksi *return* saham merupakan masalah yang sulit karena perubahan *return* tidak selalu membentuk pola yang jelas. Data pasar saham sering mengandung *noise*, yaitu perubahan yang terlihat acak dan sulit dibedakan dari pola yang benar-benar berguna untuk prediksi. Pada kondisi seperti ini, model dapat menemukan pola pada data historis, tetapi pola tersebut belum tentu tetap muncul pada data baru (Indra et al., 2025).

Penelitian pada saham LQ45 menunjukkan bahwa prediksi *return* saham Indonesia masih menghadapi sinyal yang lemah. Dalam penelitian tersebut, beberapa model *machine learning* menghasilkan nilai R-squared yang mendekati nol atau negatif, akurasi arah sekitar 49-54%, dan AUC sekitar 0,50-0,53. Hasil tersebut menunjukkan bahwa informasi historis dan indikator teknikal tidak selalu cukup kuat untuk menghasilkan prediksi *return* yang konsisten (Indra et al., 2025).

Kondisi sinyal yang lemah juga membuat model yang lebih kompleks tidak otomatis menghasilkan prediksi yang lebih baik. Model kompleks dapat mempelajari hubungan yang rumit pada data pelatihan, tetapi hubungan tersebut dapat berasal dari *noise* dan tidak berulang pada data pengujian. Oleh karena itu, performa model prediksi *return* perlu dibaca secara hati-hati, terutama ketika perbedaannya dengan metode sederhana tidak besar (Indra et al., 2025).

Kesulitan prediksi juga berkaitan dengan karakter data runtun waktu. Data runtun waktu dapat memiliki sifat *non-stationary*, yaitu pola data dapat berubah dari waktu ke waktu. Kondisi ini membuat model berbasis data tidak selalu mudah mempertahankan performanya pada periode yang berbeda. Selain itu, praktik evaluasi yang kurang tepat dapat membuat

model terlihat kompetitif, padahal hasil tersebut belum tentu bertahan pada data baru (Hewamalage et al., 2023).

Karena prediktabilitas *return* dapat rendah, evaluasi model tidak cukup dilakukan hanya dengan membandingkan beberapa model *machine learning* satu sama lain. Penelitian *forecasting* pada data dengan prediktabilitas rendah menunjukkan bahwa model yang kompleks tidak selalu mampu mengalahkan metode sederhana seperti *no-change forecast* pada data finansial yang volatil. Oleh sebab itu, penelitian prediksi *return* perlu menggunakan pembanding sederhana agar nilai tambah model dapat dibaca dengan lebih hati-hati (Beck et al., 2025).

2.2.2 Perubahan Kondisi Pasar

Selain mengandung *noise*, data pasar saham juga dapat berubah pola dari waktu ke waktu. Dalam data runtun waktu, kondisi ini berkaitan dengan *non-stationarity*, yaitu keadaan ketika pola data tidak selalu sama pada seluruh periode. Model yang terlihat baik pada data pelatihan belum tentu memiliki performa yang sama pada data baru jika pola pasar berubah. Oleh karena itu, evaluasi model prediksi *return* perlu memperhatikan urutan waktu dan perubahan kondisi pasar (Hewamalage et al., 2023).

Perubahan kondisi pasar juga dapat terlihat dari perubahan volatilitas. Volatilitas menunjukkan besar kecilnya perubahan *return* dalam suatu periode. Ketika volatilitas rendah, perubahan *return* relatif lebih kecil. Sebaliknya, ketika volatilitas tinggi, perubahan *return* cenderung lebih besar dan pasar terlihat lebih bergejolak. Pada data keuangan, volatilitas dapat berubah dari waktu ke waktu dan membentuk *volatility clustering*. Istilah ini berarti periode yang bergejolak cenderung berdekatan dengan periode bergejolak lainnya, sedangkan periode yang lebih tenang cenderung berdekatan dengan periode tenang lainnya (Adinda et al., 2026; Wiratama et al., 2025).

Kondisi tersebut membuat performa model tidak cukup dinilai dari satu nilai galat rata-rata. Sebuah model dapat menghasilkan galat yang rendah pada kondisi pasar tertentu, tetapi belum tentu memiliki galat yang sama pada kondisi pasar lainnya. Karena itu, beberapa penelitian prediksi pasar saham mulai memperhatikan kondisi atau rezim pasar dalam proses evaluasi model. Pendekatan ini membantu peneliti membaca apakah model tetap menghasilkan galat yang relatif rendah ketika kondisi pasar berubah (Pagliaro, 2026).

Dalam penelitian ini, kondisi pasar dikelompokkan berdasarkan tingkat volatilitas menjadi rendah, sedang, dan tinggi. Setelah dijelaskan sebagai kelompok volatilitas, istilah

rezim volatilitas digunakan untuk menyebut kelompok tersebut. Rezim volatilitas tidak digunakan sebagai target prediksi, tetapi sebagai cara untuk mengevaluasi performa model pada kondisi pasar yang berbeda. Dengan demikian, penelitian ini tidak menyatakan bahwa volatilitas secara langsung menyebabkan model gagal, melainkan melihat apakah galat prediksi berubah ketika model diuji pada kondisi volatilitas yang berbeda.

2.2.3 Model Markov Switching

Pengelompokan kondisi pasar dapat dilakukan dengan dua cara yang berbeda. Cara pertama adalah menetapkan batas tetap pada suatu ukuran volatilitas, misalnya membagi data menjadi tiga bagian berdasarkan nilai kuantil. Cara ini sederhana, tetapi batasnya ditentukan dari luar dan dapat berubah ketika periode datanya diganti. Cara kedua adalah menggunakan model yang memperlakukan kondisi pasar sebagai keadaan yang tidak diamati secara langsung, lalu menaksir keadaan tersebut dari data.

Model Markov Switching termasuk cara kedua. Model ini mengasumsikan bahwa data berasal dari beberapa keadaan yang berbeda, dan setiap keadaan memiliki karakteristik statistik sendiri, misalnya ragam yang berbeda. Keadaan yang sedang berlaku pada suatu waktu tidak diamati langsung, tetapi peluangnya ditaksir dari data. Perpindahan antarkeadaan mengikuti proses Markov, yaitu peluang berpindah ke keadaan berikutnya hanya bergantung pada keadaan saat ini (Hamilton, 1989).

Karena keadaan ditaksir dari data, batas antarkeadaan tidak ditetapkan lebih dahulu oleh peneliti melainkan mengikuti pola ragam yang ditemukan pada data pelatihan. Sifat ini membuat pengelompokan kondisi pasar tidak bergantung pada pemilihan nilai ambang tertentu.

Pendekatan berbasis keadaan pasar sudah digunakan pada penelitian prediksi saham. Salah satu penelitian menggunakan model keadaan tersembunyi yang diperbarui secara bergulir dengan hanya memakai data masa lalu agar informasi masa depan tidak ikut terpakai (Pagliaro, 2026). Meskipun demikian, manfaat memasukkan informasi rezim ke dalam model belum tentu terbukti. Penelitian yang membandingkan model tanpa informasi rezim, model dengan rezim sebagai fitur, dan model terpisah untuk setiap rezim justru menemukan bahwa model tanpa informasi rezim memberikan hasil terbaik (Gomez Grau, 2025).

Dalam penelitian ini, Markov Switching digunakan hanya untuk membentuk kelompok evaluasi, bukan sebagai model prediksi maupun sebagai fitur. Rincian jumlah keadaan, data

yang dipakai untuk estimasi, serta cara penetapan label rendah, sedang, dan tinggi dijelaskan pada Bab 3.

2.3 *Machine learning* untuk Prediksi *Return* Saham

Machine learning dapat digunakan untuk mempelajari pola dari data historis dan menghasilkan prediksi pada data baru. Namun, pada data pasar saham, model tetap perlu diuji dengan hati-hati karena pola historis tidak selalu kuat dan dapat berubah dari waktu ke waktu. Oleh karena itu, penelitian ini tidak hanya menggunakan satu model, tetapi membandingkan beberapa model dengan karakteristik yang berbeda (Hewamalage et al., 2023).

2.3.1 Ridge Regression

Ridge Regression merupakan model regresi linear yang menambahkan regularisasi L2. Regresi linear berusaha mempelajari hubungan antara fitur masukan dan target prediksi, sedangkan regularisasi L2 memberi penalti pada koefisien model agar nilainya tidak menjadi terlalu besar. Koefisien dapat dipahami sebagai bobot yang menunjukkan seberapa besar pengaruh suatu fitur dalam model. Dengan adanya penalti tersebut, Ridge Regression membantu model agar tidak terlalu mengikuti pola pada data pelatihan (Indra et al., 2025; Saputra et al., 2025).

Regularisasi L2 berguna ketika fitur yang digunakan saling berkaitan. Dalam prediksi berbasis indikator teknikal, beberapa fitur dapat membawa informasi yang mirip. Kondisi ini disebut multikolinearitas, yaitu keadaan ketika beberapa fitur memiliki hubungan yang kuat satu sama lain. Jika tidak dikendalikan, model dapat memberi bobot yang terlalu besar pada fitur tertentu dan berisiko mengalami *overfitting*. *Overfitting* berarti model terlalu menyesuaikan diri dengan data pelatihan, tetapi tidak bekerja sebaik itu pada data baru (Indra et al., 2025).

Kekuatan regularisasi pada Ridge Regression diatur oleh parameter alpha. Alpha yang lebih besar memberi penalti lebih kuat pada koefisien, sedangkan alpha yang lebih kecil memberi penalti lebih lemah. Dalam penelitian ini, alpha disesuaikan pada proses penalaan parameter agar model dapat diuji dengan pengaturan yang sesuai pada setiap data pelatihan. Penjelasan teknis mengenai proses penalaan parameter dan rumus objektif Ridge Regression disampaikan pada Bab 3.

Ridge Regression relevan digunakan dalam penelitian ini karena model tersebut telah digunakan pada prediksi *return* saham Indonesia. Penelitian pada saham perbankan Indonesia

menggunakan Ridge Regression dengan fitur teknikal seperti *moving average* dan *rolling standard deviation* untuk memprediksi *return* saham. Selain itu, penelitian pada saham LQ45 juga menggunakan Ridge Regression bersama model lain, seperti Linear Regression, Random Forest, dan XGBoost, untuk memprediksi *return* saham berbasis indikator teknikal (Hewamalage et al., 2023; Saputra et al., 2025).

2.3.2 Random Forest

Random Forest merupakan model *ensemble* berbasis banyak pohon keputusan. *Ensemble* berarti gabungan dari beberapa model yang digunakan bersama untuk menghasilkan prediksi. Pohon keputusan sendiri dapat dipahami sebagai model yang membuat keputusan melalui serangkaian pemisahan kondisi pada fitur. Jika satu pohon keputusan hanya menghasilkan prediksi berdasarkan satu struktur pohon, Random Forest menggabungkan banyak pohon agar prediksi tidak terlalu bergantung pada satu pohon saja (Breiman, 2001).

Dalam Random Forest, pohon-pohon dibentuk dengan unsur acak. Unsur acak tersebut membuat setiap pohon tidak sama persis, sehingga hasil antar pohon dapat saling melengkapi. Gagasan ini penting karena Random Forest tidak hanya mengandalkan kekuatan satu pohon, tetapi juga mengurangi ketergantungan antarpohon. Breiman menjelaskan bahwa Random Forest terdiri dari kumpulan *tree predictors* yang masing-masing bergantung pada *random vector*, dan konsep ini juga dapat digunakan untuk masalah regresi (Breiman, 2001).

Random Forest relevan untuk prediksi *return* karena model ini dapat menangkap hubungan nonlinier antara fitur dan target. Hubungan nonlinier berarti perubahan target tidak selalu mengikuti pola garis lurus dari fitur yang digunakan. Dalam konteks prediksi saham, hubungan antara indikator teknikal dan *return* dapat bersifat tidak sederhana. Penelitian pada saham LQ45 menggunakan Random Forest bersama Linear Regression, Ridge Regression, dan XGBoost untuk memprediksi *return* saham berbasis indikator teknikal (Indra et al., 2025).

Dalam penelitian ini, Random Forest digunakan sebagai salah satu kandidat model *tree ensemble*. Model ini tidak diperlakukan sebagai model utama sejak awal, tetapi diuji bersama Ridge Regression, LightGBM, dan XGBoost pada dataset dan skema evaluasi yang sama. Dengan demikian, performa Random Forest tetap ditentukan berdasarkan hasil evaluasi penelitian, bukan berdasarkan asumsi bahwa model ini selalu lebih baik daripada model linear atau model *boosting*.

2.3.3 LightGBM

LightGBM merupakan model *Gradient Boosting Decision Tree*. *Gradient Boosting Decision Tree* adalah model *ensemble* yang menggunakan kumpulan pohon keputusan. Berbeda dari Random Forest yang membangun banyak pohon dengan unsur acak, *boosting* membangun model secara bertahap. Pada proses *boosting*, model berikutnya berusaha memperbaiki kesalahan dari model sebelumnya, sehingga prediksi akhir dibentuk dari gabungan beberapa model yang disusun secara berurutan (Ke et al., 2017).

LightGBM dikembangkan untuk membuat proses pelatihan *Gradient Boosting Decision Tree* menjadi lebih efisien. Efisiensi ini penting ketika data memiliki banyak fitur atau jumlah observasi yang besar. Paper utama LightGBM menjelaskan bahwa pelatihan GBDT dapat menjadi mahal secara komputasi karena model perlu mencari titik pemisahan terbaik pada banyak fitur. Untuk mengatasi masalah tersebut, LightGBM menggunakan mekanisme seperti *Gradient-based One-Side Sampling* (GOSS) dan *Exclusive Feature Bundling* (EFB) (Ke et al., 2017)

Secara sederhana, GOSS membantu model mempertahankan data yang lebih berpengaruh dalam proses pelatihan, sedangkan EFB membantu menggabungkan fitur yang jarang aktif secara bersamaan agar jumlah fitur efektif menjadi lebih ringkas. Dengan mekanisme tersebut, LightGBM dirancang agar pelatihan model *boosting* dapat berjalan lebih efisien (Ke et al., 2017)

LightGBM juga dikenal menggunakan strategi pertumbuhan pohon secara *leaf-wise*. *Leaf-wise growth* berarti pohon dikembangkan dengan memilih bagian daun yang dianggap paling berpengaruh terhadap penurunan galat model. Strategi ini berkaitan dengan cara LightGBM membentuk pohon keputusan, sehingga pengaturannya tetap perlu dikendalikan melalui penalaan parameter agar model tidak terlalu menyesuaikan diri dengan data pelatihan (Ke et al., 2017).

Dalam konteks data finansial, LightGBM relevan karena dapat digunakan pada data tabular, yaitu data yang tersusun dalam bentuk baris dan kolom seperti dataset fitur teknikal, variabel makroekonomi, dan target *return*. Penelitian terdahulu menggunakan LightGBM dengan fitur teknikal, makro, dan fitur pasar lain dalam kerangka prediksi pasar saham berbasis validasi *walk-forward*. Hal ini menunjukkan bahwa LightGBM dapat dipakai sebagai salah (Pagliaro, 2026).

2.3.4 XGBoost

XGBoost merupakan model *gradient tree boosting* berbasis kumpulan pohon keputusan. *Gradient tree boosting* dapat dipahami sebagai pendekatan yang membangun model secara bertahap. Pada setiap tahap, pohon keputusan baru ditambahkan untuk memperbaiki kesalahan prediksi dari model sebelumnya. Dengan cara ini, prediksi akhir XGBoost diperoleh dari gabungan beberapa pohon yang bekerja secara berurutan (Chen & Guestrin, 2016).

XGBoost termasuk model *tree ensemble*. *Tree ensemble* berarti model menggunakan lebih dari satu pohon keputusan untuk menghasilkan prediksi. Berbeda dari model linear yang membentuk hubungan dalam bentuk garis lurus, model berbasis pohon dapat membagi data berdasarkan kondisi pada fitur. Karena itu, XGBoost dapat digunakan untuk mempelajari hubungan nonlinier antara fitur dan target (Chen & Guestrin, 2016).

Salah satu karakteristik penting XGBoost adalah adanya *regularized learning objective*. Regularisasi berarti mekanisme untuk membatasi kompleksitas model. Dalam XGBoost, regularisasi digunakan agar model tidak membentuk struktur pohon yang terlalu kompleks dan tidak terlalu menyesuaikan diri dengan data pelatihan. Kondisi ketika model terlalu mengikuti data pelatihan tetapi kurang baik pada data baru disebut *overfitting* (Chen & Guestrin, 2016).

Selain regularisasi, XGBoost juga memiliki beberapa mekanisme yang mendukung efisiensi komputasi. Paper utama XGBoost menjelaskan adanya teknik seperti *sparsity-aware algorithm*, *weighted quantile sketch*, serta optimasi sistem untuk membuat proses pelatihan lebih efisien. Dalam penelitian ini, detail matematis dan sistem komputasi tersebut tidak dibahas panjang karena fokus penelitian adalah evaluasi performa model, bukan pengembangan algoritma XGBoost (Chen & Guestrin, 2016).

XGBoost relevan digunakan dalam penelitian ini karena model tersebut pernah digunakan dalam penelitian prediksi *return* saham Indonesia berbasis indikator teknikal. Penelitian pada saham LQ45 membandingkan beberapa model, termasuk Ridge Regression, Random Forest, dan XGBoost, untuk memprediksi *return* saham. Dalam konteks tersebut, XGBoost digunakan sebagai salah satu model berbasis pohon yang dapat menangkap hubungan nonlinier antara fitur teknikal dan *return* (Indra et al., 2025).

Dalam penelitian ini, XGBoost digunakan sebagai salah satu kandidat model *boosting*. Model ini dipilih untuk dibandingkan dengan Ridge Regression, Random Forest, dan LightGBM pada dataset, target, dan skema evaluasi yang sama. Namun, penggunaan XGBoost tidak berarti bahwa model ini diasumsikan selalu lebih baik daripada model lain. Kesimpulan mengenai performa XGBoost tetap ditentukan berdasarkan hasil evaluasi penelitian.

2.4 Fitur Prediksi yang Digunakan

Fitur prediksi adalah variabel masukan yang digunakan model untuk menghasilkan prediksi. Dalam penelitian ini, fitur prediksi dikelompokkan menjadi fitur teknikal, variabel makroekonomi, dan fitur volatilitas berbasis GARCH. Subbab ini menjelaskan dasar penggunaan masing-masing kelompok fitur. Detail pembentukan dataset dan tahapan pengolahannya dijelaskan pada Bab 3.

2.4.1 Fitur Teknikal

Fitur teknikal adalah variabel masukan model yang dibentuk dari data historis harga dan volume. Pada penelitian prediksi *return* saham, fitur teknikal digunakan untuk memberi informasi tentang perubahan harga, tren, momentum, volatilitas, dan aktivitas volume. Dengan menggunakan fitur tersebut, model tidak hanya menerima nilai harga pada satu waktu, tetapi juga ringkasan informasi historis yang sudah terjadi sebelum prediksi dibuat (Indra et al., 2025).

Dalam penelitian ini, beberapa indikator teknikal digunakan sebagai fitur. Fitur historis seperti *return* lag digunakan untuk membawa informasi dari beberapa sesi sebelumnya. *Exponential Moving Average* (EMA) merangkum kecenderungan harga dengan memberi bobot yang menurun secara bertahap pada pengamatan yang semakin lama, sehingga EMA bereaksi lebih cepat terhadap perubahan terbaru dibandingkan rata-rata bergerak sederhana. *Moving Average Convergence Divergence* (MACD) dibentuk dari selisih dua EMA dengan panjang periode berbeda, sehingga nilainya membaca apakah kecenderungan jangka pendek sedang menjauh atau mendekati kecenderungan jangka panjang. *Average True Range* (ATR) mengukur rentang pergerakan harga dengan memperhitungkan selisih tertinggi dan terendah beserta lompatan harga antarsesi, sehingga ATR membaca besar pergerakan tanpa memandang arahnya. *Relative Strength Index* (RSI) membandingkan besar kenaikan dan penurunan harga pada suatu jendela waktu dan dinyatakan pada skala nol sampai seratus, sehingga RSI membaca kekuatan relatif pergerakan harga. *On-Balance Volume* (OBV) menjumlahkan volume perdagangan dengan tanda mengikuti arah perubahan harga, sehingga OBV membaca apakah aktivitas volume searah dengan pergerakan harga. Selain kelima indikator tersebut, penelitian ini juga menggunakan momentum, *rolling volatility*, *Bollinger Band*, *stochastic oscillator*, dan rasio volume (Indra et al., 2025; Lutey, 2022; Muhammad, 2024; Tsang & Chong, 2009).

Kelima indikator tersebut mewakili empat jenis informasi yang berbeda, yaitu kecenderungan harga, kekuatan dan arah pergerakan, besar fluktuasi, serta aktivitas volume. Pengelompokan tersebut digunakan agar fitur teknikal tidak hanya memuat pengulangan informasi harga, tetapi juga memuat ringkasan perilaku pasar dari sudut pandang yang berbeda. Rincian fitur yang dibentuk pada setiap kelompok beserta panjang jendela waktunya dijelaskan pada Bab 3.

Penggunaan fitur teknikal tidak berarti setiap indikator pasti meningkatkan performa model. Literatur menunjukkan bahwa manfaat indikator teknikal dapat berbeda tergantung pada data, periode, dan model yang digunakan. Oleh karena itu, dalam penelitian ini fitur teknikal diperlakukan sebagai sumber informasi yang diuji melalui skema evaluasi, bukan sebagai jaminan bahwa model akan menghasilkan galat yang lebih rendah (A. Deep, Monico, Shirvani, Rachev, & Fabozzi, 2024).

Dalam penelitian ini, fitur teknikal menjadi dasar pembentukan dataset. Informasi teknikal digunakan sebagai kelompok fitur utama, sedangkan variabel makroekonomi dan fitur volatilitas ditambahkan pada rancangan dataset berikutnya untuk menguji apakah informasi tambahan berkaitan dengan perubahan galat model.

2.4.2 Variabel Makroekonomi

Pergerakan pasar saham tidak hanya dapat dibaca dari harga dan volume historis. Pasar saham juga dapat berkaitan dengan kondisi ekonomi yang lebih luas, baik dari dalam negeri maupun dari luar negeri. Dalam konteks IHSG, penelitian terdahulu membahas hubungan variabel makroekonomi seperti suku bunga, inflasi, dan nilai tukar USD/IDR dengan pergerakan indeks saham. Hal ini menunjukkan bahwa variabel makroekonomi dapat dipertimbangkan sebagai informasi tambahan dalam pemodelan pasar saham (Marpaung & Pangestuti, 2024).

Dalam penelitian ini, variabel makroekonomi digunakan sebagai fitur tambahan di luar fitur teknikal. Variabel domestik yang digunakan meliputi nilai tukar USD/IDR, suku bunga Indonesia, dan inflasi Indonesia. Nilai tukar USD/IDR digunakan untuk membaca konteks hubungan rupiah terhadap dolar AS. Suku bunga digunakan untuk menggambarkan konteks kebijakan moneter dan biaya dana dalam perekonomian. Inflasi digunakan untuk membaca perubahan tingkat harga secara umum. Ketiga variabel tersebut dipakai sebagai informasi ekonomi domestik yang dapat berkaitan dengan kondisi pasar saham Indonesia (Marpaung & Pangestuti, 2024).

Selain variabel domestik, penelitian ini juga menggunakan variabel global. Variabel global yang digunakan meliputi S&P 500, US 10-Year Treasury Yield, dan Dollar Index. S&P 500 digunakan sebagai indikator pasar saham Amerika Serikat yang dapat memberi gambaran kondisi pasar global. US 10-Year Treasury Yield digunakan sebagai indikator konteks suku bunga global, sedangkan Dollar Index digunakan untuk membaca kekuatan dolar AS secara umum. Penelitian terdahulu pada prediksi saham berbasis LightGBM juga menggunakan fitur makro dan *cross-asset* seperti S&P 500 ETF, Treasury, dan US Dollar Index untuk memberi konteks pasar yang lebih luas (Pagliaro, 2026).

Penggunaan variabel makroekonomi tidak berarti bahwa variabel tersebut pasti meningkatkan performa model. Hubungan antara variabel makro dan pasar saham dapat berbeda antarperiode, pasar, dan konteks penelitian. Oleh karena itu, variabel makro dalam penelitian ini diperlakukan sebagai informasi tambahan yang diuji melalui evaluasi model, bukan sebagai bukti bahwa faktor makro secara langsung menyebabkan perubahan *return* IHSG (Marpaung & Pangestuti, 2024; Situmorang & Kaluge, 2025).

Dalam penelitian ini, variabel makroekonomi digunakan sebagai informasi tambahan di luar fitur teknikal. Variabel domestik digunakan untuk membentuk dataset dengan informasi ekonomi Indonesia, sedangkan variabel global digunakan sebagai bagian dari dataset yang memuat informasi makro inti. Detail pembentukan dataset dijelaskan pada Bab 3.

2.4.3 Fitur Volatilitas Berbasis GARCH

Volatilitas dalam penelitian ini dibaca sebagai besar kecilnya perubahan *return* dalam suatu periode. Pada data keuangan, volatilitas tidak selalu berada pada tingkat yang sama dari waktu ke waktu. Ada periode ketika perubahan *return* relatif kecil dan pasar terlihat lebih tenang, tetapi ada juga periode ketika perubahan *return* lebih besar dan pasar terlihat lebih bergejolak. Kondisi seperti ini berkaitan dengan *volatility clustering*, yaitu keadaan ketika periode volatilitas tinggi cenderung berdekatan dengan periode volatilitas tinggi lainnya, sedangkan periode volatilitas rendah cenderung berdekatan dengan periode yang lebih tenang (Adinda et al., 2026; Wiratama et al., 2025).

Perubahan volatilitas tersebut juga berkaitan dengan heteroskedastisitas. Secara sederhana, heteroskedastisitas berarti tingkat variasi atau fluktuasi data tidak konstan sepanjang waktu. Pada data pasar saham, kondisi ini membuat model perlu memperhatikan bahwa besar kecilnya perubahan *return* dapat berbeda antarperiode. Penelitian terdahulu pada IHSG

menunjukkan bahwa model ARCH atau GARCH digunakan untuk menangani kondisi ketika varians atau volatilitas berubah dari waktu ke waktu (Adinda et al., 2026).

GARCH atau *Generalized Autoregressive Conditional Heteroskedasticity* merupakan pendekatan yang digunakan untuk mengestimasi volatilitas pada data keuangan. GARCH dapat menangkap pola volatilitas yang berubah dari waktu ke waktu dan digunakan dalam penelitian volatilitas indeks saham. Dalam penelitian tentang volatilitas IHSG, keluarga GARCH digunakan untuk menangkap karakteristik *volatility clustering* dan menghasilkan estimasi volatilitas bersyarat (Wiratama et al., 2025).

Dalam penelitian ini, GARCH tidak digunakan sebagai model utama dan tidak mengubah target penelitian menjadi prediksi volatilitas. Fitur GARCH digunakan sebagai informasi volatilitas tambahan pada Dataset C.

Penggunaan fitur volatilitas berbasis GARCH tidak berarti fitur tersebut pasti meningkatkan performa model. Fitur ini digunakan untuk menguji apakah informasi volatilitas tambahan berkaitan dengan perubahan galat model. Dengan demikian, peran GARCH dalam penelitian ini bersifat sebagai fitur pendukung pada Dataset C, bukan sebagai kesimpulan bahwa estimasi volatilitas selalu membuat prediksi *return* menjadi lebih baik.

2.4.4 Ablation Study

Ketika sebuah model menggunakan beberapa kelompok fitur sekaligus, hasil evaluasi keseluruhan tidak menunjukkan kelompok mana yang benar-benar berpengaruh. Nilai galat yang rendah dapat berasal dari satu kelompok fitur saja, dari gabungan beberapa kelompok, atau bahkan tetap rendah meskipun salah satu kelompok dihilangkan.

Ablation study adalah cara untuk menilai kontribusi setiap komponen dengan membandingkan hasil sebelum dan sesudah komponen tersebut ditambahkan atau dihilangkan, sedangkan seluruh kondisi lain dibuat sama. Karena model, data pengujian, dan prosedur evaluasi tidak diubah, perbedaan hasil yang muncul dapat dikaitkan dengan keberadaan komponen yang diuji. Penelitian prediksi saham menggunakan cara ini untuk menilai kelompok fitur mana yang memberi kontribusi prediktif paling besar (Pagliaro, 2026).

Hasil *ablation study* dibaca sebagai kontribusi prediktif pada sampel evaluasi yang digunakan, bukan sebagai bukti hubungan sebab-akibat antara kelompok fitur dan *return*. Kontribusi yang terlihat pada satu periode atau satu susunan model belum tentu sama pada periode atau susunan yang lain.

Dalam penelitian ini, *ablation study* digunakan untuk menilai perubahan hasil ketika indikator makroekonomi domestik, indikator pasar global, dan fitur volatilitas ditambahkan secara bertahap. Tahapan perbandingan dan cara pengukurannya dijelaskan pada Bab 3.

2.5 Metode Pembanding Sederhana

Evaluasi model prediksi tidak cukup dilakukan dengan membandingkan model *machine learning* satu sama lain. Model juga perlu dibandingkan dengan metode pembanding sederhana agar nilai tambah model dapat dibaca dengan lebih hati-hati. Dalam *forecasting*, pembanding sederhana penting karena model yang lebih kompleks belum tentu menghasilkan prediksi yang lebih baik, terutama pada data yang memiliki prediktabilitas rendah atau banyak fluktuasi acak. Oleh karena itu, subbab ini membahas dua metode pembanding sederhana yang digunakan dalam penelitian ini, yaitu *no-change forecast* dan *persistence forecast* (Beck et al., 2025; Hewamalage et al., 2023).

2.5.1 No-change Forecast

No-change forecast adalah metode pembanding sederhana yang mengasumsikan tidak ada perubahan dari titik prediksi. Dalam ruang harga, asumsi ini berarti nilai indeks pada periode berikutnya dianggap sama dengan nilai indeks saat prediksi dibuat. Karena penelitian ini memprediksi *return*, maka *no-change forecast* digunakan sebagai prediksi *return* sebesar 0%. Sebagai contoh, jika IHSG berada pada nilai 7.000 saat prediksi dibuat, maka *no-change forecast* menganggap nilai IHSG pada akhir horizon prediksi tetap setara dengan 7.000. Dengan demikian, *return* yang diprediksi adalah 0%. Berbeda dari *persistence forecast* yang menggunakan *return* historis terbaru sebagai acuan, *no-change forecast* tidak mengambil *return* periode sebelumnya sebagai nilai prediksi.

Metode ini terlihat sangat sederhana, tetapi tetap penting dalam evaluasi model prediksi. Dalam *forecasting*, metode sederhana seperti *naive forecast* atau *no-change model* sering digunakan sebagai pembanding awal untuk menilai apakah model yang lebih kompleks memberikan nilai tambah. Literatur evaluasi *forecasting* menjelaskan bahwa model baru perlu dibandingkan dengan *benchmark* yang tepat, terutama *benchmark* sederhana, agar performanya tidak hanya terlihat baik ketika dibandingkan dengan model kompleks lain (Hewamalage et al., 2023).

No-change forecast juga berkaitan dengan *random walk benchmark*. Secara umum, *random walk* menggambarkan keadaan ketika nilai berikutnya bergerak dari nilai saat ini

dengan tambahan perubahan acak. Jika data tidak memiliki pola prediktif yang kuat, prediksi sederhana yang mengacu pada kondisi terakhir dapat menjadi pembanding yang sulit dikalahkan. Pada data *time series* dengan prediktabilitas rendah, terutama data finansial yang volatil, penelitian terdahulu menunjukkan bahwa model yang kompleks tidak selalu mampu mengungguli metode pembanding sederhana seperti *no-change forecast* (Beck et al., 2025).

Dalam penelitian prediksi *return* saham, penggunaan *no-change forecast* membantu menilai apakah model *machine learning* menghasilkan galat yang lebih rendah daripada asumsi paling sederhana, yaitu tidak ada perubahan *return*. Oleh karena itu, *no-change forecast* tidak digunakan sebagai model utama, melainkan sebagai titik acuan untuk membaca nilai tambah prediksi model. Jika model menghasilkan galat yang lebih rendah daripada *no-change forecast*, hasil tersebut dapat dibaca sebagai perbaikan terhadap pembanding sederhana. Sebaliknya, jika model tidak mampu mengunggulinya, hasil model perlu ditafsirkan lebih hati-hati

2.5.2 Persistence Forecast

Persistence forecast adalah metode pembanding sederhana yang menggunakan nilai historis terbaru sebagai prediksi masa depan. Dalam konteks *forecasting*, pendekatan ini berangkat dari aturan sederhana bahwa nilai yang baru saja terjadi dapat digunakan sebagai acuan untuk periode berikutnya. Literatur *forecasting* menjelaskan bahwa *persistence model* termasuk salah satu bentuk *benchmark* sederhana yang memakai observasi terakhir sebagai prediksi (Hewamalage et al., 2023).

Dalam penelitian ini, target yang diprediksi adalah *return* IHSG 20 sesi perdagangan ke depan. Oleh karena itu, *persistence forecast* menggunakan *return* IHSG pada 20 sesi perdagangan terakhir sebagai prediksi *return* 20 sesi perdagangan berikutnya. Sebagai contoh, jika *return* IHSG selama 20 sesi terakhir sebesar 3%, maka *persistence forecast* memprediksi *return* 20 sesi berikutnya juga sebesar 3%. Dengan cara ini, *persistence forecast* mewakili asumsi sederhana bahwa kondisi *return* terbaru berlanjut ke periode berikutnya.

Persistence forecast berbeda dari *no-change forecast*. *No-change forecast* memprediksi *return* sebesar 0%, sedangkan *persistence forecast* mengikuti *return* historis terbaru. Perbedaan ini penting karena kedua metode tersebut memberikan titik pembanding yang berbeda. *No-change forecast* mewakili asumsi tidak ada perubahan, sedangkan *persistence forecast* mewakili aturan bahwa pola *return* terbaru berlanjut ke periode berikutnya.

Metode ini digunakan sebagai pembanding karena model *machine learning* perlu menunjukkan apakah prediksinya lebih baik daripada aturan sederhana yang hanya

meneruskan informasi terakhir. Pada penelitian prediksi *return*, *persistence forecast* juga digunakan sebagai pembanding terhadap model yang lebih kompleks, misalnya dengan memprediksi *return* berikutnya berdasarkan *return* terakhir yang diamati (Spulbăr & Ene, 2025). Dengan demikian, *persistence forecast* tidak digunakan sebagai model utama, tetapi sebagai acuan untuk menilai nilai tambah prediksi model *machine learning*.

2.6 Validasi Model pada Data *Time Series*

Data IHSG merupakan data *time series* karena setiap observasi memiliki urutan tanggal. Pada data seperti ini, evaluasi model perlu mempertahankan urutan waktu agar proses pengujian menyerupai kondisi prediksi yang sebenarnya. Model dilatih menggunakan data yang sudah terjadi, lalu diuji pada periode setelahnya. Oleh karena itu, pembagian data pada penelitian prediksi *time series* tidak dilakukan secara acak karena dapat mencampur informasi masa lalu dan masa depan (Hewamalage et al., 2023).

2.6.1 Chronological Holdout

Chronological holdout adalah cara membagi data pelatihan dan data pengujian berdasarkan urutan waktu. Data yang lebih lama digunakan sebagai data pelatihan, sedangkan data yang lebih baru digunakan sebagai data pengujian. Pada konteks prediksi IHSG, cara ini penting karena model seharusnya belajar dari data historis terlebih dahulu sebelum diuji pada periode setelahnya. Pendekatan ini sejalan dengan prinsip *time-ordered split* pada data saham, yaitu pembagian data yang mempertahankan sifat temporal data dan mengurangi risiko tercampurnya informasi antarperiode (G. Deep et al., 2025).

Dalam penelitian ini, *chronological holdout* digunakan sebagai evaluasi awal untuk melihat performa model pada satu pembagian waktu. Evaluasi ini membantu memberikan gambaran awal tentang kemampuan model pada data yang tidak digunakan saat pelatihan, atau *out-of-sample*. *Out-of-sample* berarti model diuji pada data baru yang terpisah dari data pelatihan. Dalam *forecasting*, evaluasi pada *validation set* atau *test set* yang terpisah dari data pelatihan juga dikenal sebagai *out-of-sample evaluation* (Hewamalage et al., 2023). Namun, karena satu pembagian waktu dapat dipengaruhi oleh periode yang dipilih, evaluasi utama dalam penelitian ini tetap menggunakan *walk-forward validation* yang dijelaskan pada subbab berikutnya.

2.6.2 Walk-forward Validation

Walk-forward validation adalah cara mengevaluasi model pada data *time series* secara berulang. Pada setiap tahap, model dilatih menggunakan data yang berada sebelum periode uji, lalu diuji pada periode setelahnya. Setelah satu tahap selesai, proses bergerak maju mengikuti urutan waktu. Pendekatan ini sesuai untuk data runtun waktu karena mempertahankan hubungan temporal data (Wahyuddin et al., 2025). Konsep ini juga sejalan dengan *rolling origin evaluation*, yaitu evaluasi ketika titik asal prediksi bergerak maju sehingga menghasilkan beberapa periode pengujian *out-of-sample* (Hewamalage et al., 2023).

Dalam penelitian ini, *walk-forward validation* digunakan dalam bentuk *expanding window*. *Expanding window* berarti data pelatihan bertambah dari *fold* ke *fold*. *Fold* dapat dipahami sebagai satu tahap pelatihan dan pengujian. Pada *fold* berikutnya, model dapat menggunakan lebih banyak data historis dibandingkan *fold* sebelumnya. Literatur *forecasting* menjelaskan bahwa *rolling origin evaluation* dapat dilakukan dengan *expanding window*, yaitu data pelatihan terus bertambah seiring bertambahnya waktu (Hewamalage et al., 2023).

Pendekatan ini membantu mengevaluasi model pada beberapa periode *out-of-sample*. Dengan demikian, hasil evaluasi tidak hanya bergantung pada satu pembagian waktu seperti pada *chronological holdout*. Satu pembagian waktu dapat dipengaruhi oleh karakter periode uji tertentu, sehingga evaluasi berulang dapat memberi gambaran yang lebih luas tentang perubahan performa model dari periode ke periode. Literatur evaluasi *forecasting* menjelaskan bahwa *fixed origin setup* dapat menghasilkan galat yang dipengaruhi pola khusus pada satu horizon uji, sehingga evaluasi multiperiode dapat digunakan sebagai pembanding yang lebih informatif (Hewamalage et al., 2023).

Walk-forward validation penting dalam penelitian ini karena fokus penelitian tidak hanya mencari model dengan galat rata-rata paling rendah, tetapi juga melihat perubahan performa model dari waktu ke waktu. Dengan mengevaluasi model pada beberapa periode uji, penelitian dapat membaca apakah performa model relatif konsisten atau berubah ketika periode pengujian berganti. Namun, penggunaan *walk-forward validation* tidak berarti model dijamin stabil pada masa depan. Metode ini hanya membantu membuat evaluasi lebih sesuai dengan karakter data runtun waktu dan lebih kaya dibandingkan satu pembagian data.

2.6.3 Data Leakage dan Look-ahead Bias

Data leakage adalah kondisi ketika informasi yang seharusnya belum tersedia pada saat prediksi ikut masuk ke proses pelatihan, validasi, atau evaluasi model. Pada data *time series*,

risiko ini perlu diperhatikan karena model seharusnya hanya belajar dari data yang sudah terjadi sebelum periode uji. Sebagai contoh, model tidak boleh memakai informasi dari periode uji ketika sedang mempelajari pola dari data pelatihan. Jika hal tersebut terjadi, hasil evaluasi dapat terlihat terlalu baik karena model secara tidak langsung memperoleh informasi yang tidak tersedia dalam kondisi prediksi yang sebenarnya (Hewamalage et al., 2023).

Look-ahead bias berkaitan erat dengan *data leakage*. *Look-ahead bias* terjadi ketika model menggunakan informasi dari masa depan untuk membuat prediksi pada periode sebelumnya atau periode saat ini. Dalam konteks prediksi IHSG, model seharusnya membuat prediksi berdasarkan informasi yang tersedia sampai waktu prediksi dibuat. Jika fitur, target, atau proses evaluasi memasukkan informasi setelah waktu prediksi, maka evaluasi tidak lagi mencerminkan situasi prediksi yang realistis. Penelitian prediksi finansial menekankan pentingnya menjaga keselarasan waktu agar model tidak menggunakan informasi masa depan dalam proses prediksi (Liu, 2026; Pagliaro, 2026).

Pencegahan *data leakage* tidak hanya dilakukan melalui pembagian data berdasarkan waktu. Tahap preprocessing dan transformasi fitur juga perlu dilakukan dengan hati-hati. Prapemrosesan adalah tahap penyiapan data sebelum model dilatih, sedangkan transformasi fitur adalah perubahan nilai fitur agar lebih siap digunakan oleh model. Parameter transformasi seperti rata-rata dan simpangan baku untuk standardisasi sebaiknya dihitung dari data pelatihan, lalu diterapkan pada data validasi atau data uji. Dalam rancangan metodologi penelitian ini, prinsip yang sama juga digunakan pada batas *winsorization*. Dengan cara ini, data validasi dan data uji tidak ikut memengaruhi proses pelatihan model (Liu, 2026; Pagliaro, 2026).

Chronological holdout dan *walk-forward validation* membantu menjaga urutan waktu karena data pelatihan ditempatkan sebelum data pengujian. Namun, menjaga urutan waktu saja belum cukup apabila proses pembentukan fitur, prapemrosesan, atau penalaan parameter masih memakai informasi dari periode validasi atau uji. Oleh karena itu, rancangan evaluasi pada data runtun waktu perlu memastikan bahwa setiap tahapan hanya menggunakan informasi yang tersedia pada waktu yang sesuai. Prinsip ini penting agar hasil evaluasi tidak terlalu optimistis dan lebih mencerminkan kemampuan model pada data baru.

Dalam penelitian ini, prosedur evaluasi dirancang untuk mengurangi risiko *data leakage* dan *look-ahead bias*. Fitur dibentuk berdasarkan informasi historis, pembagian data mengikuti urutan waktu, dan prapemrosesan dilakukan berdasarkan data pelatihan. Namun, penggunaan prosedur tersebut tidak berarti bahwa seluruh risiko bias dapat dinyatakan hilang sepenuhnya.

Prosedur tersebut lebih tepat dipahami sebagai upaya untuk membatasi penggunaan informasi masa depan agar evaluasi model menjadi lebih hati-hati dan sesuai dengan karakter data runtun waktu.

2.7 Metrik Evaluasi Model

2.7.1 RMSE

RMSE atau *Root Mean Squared Error* adalah metrik galat regresi yang digunakan untuk mengukur jarak antara nilai aktual dan nilai prediksi. Galat dalam konteks ini berarti selisih antara *return* aktual dan *return* yang diprediksi oleh model. RMSE menghitung galat dengan cara menguadratkan selisih antara nilai aktual dan nilai prediksi, menghitung rata-ratanya, lalu mengambil akar dari rata-rata tersebut. Dalam evaluasi *forecasting*, RMSE termasuk ukuran galat yang umum digunakan untuk menilai performa prediksi pada data bernilai kontinu (Hewamalage et al., 2023).

Karena galat dikuadratkan terlebih dahulu, RMSE menjadi lebih sensitif terhadap galat yang besar. Artinya, kesalahan prediksi yang besar akan memberi pengaruh lebih kuat terhadap nilai RMSE dibandingkan kesalahan yang kecil. Karakteristik ini membuat RMSE berguna ketika penelitian ingin memberi perhatian lebih pada kesalahan prediksi yang besar, bukan hanya melihat rata-rata kesalahan secara umum. Dalam konteks prediksi *return*, hal ini penting karena galat yang besar menunjukkan selisih yang lebih besar antara *return* aktual dan *return* hasil prediksi.

Dalam penelitian ini, RMSE digunakan sebagai metrik utama karena target prediksi berupa *return* IHSG yang berbentuk nilai kontinu. Nilai kontinu berarti nilai prediksi tidak berupa kelas seperti naik atau turun, tetapi berupa angka *return*. Nilai RMSE yang lebih kecil menunjukkan bahwa prediksi model lebih dekat dengan *return* aktual. RMSE juga digunakan dalam penelitian prediksi saham dan *equity premium* sebagai metrik untuk menilai besar galat prediksi. Oleh karena itu, RMSE sesuai digunakan sebagai metrik utama dalam penelitian ini karena target yang diprediksi berupa *return* kontinu (Cao, 2025).

2.7.2 MAE

MAE atau *Mean Absolute Error* adalah metrik galat regresi yang menghitung rata-rata nilai absolut dari selisih antara nilai aktual dan nilai prediksi. Galat absolut berarti besar kesalahan dihitung tanpa memperhatikan arah kesalahannya. Dengan menggunakan nilai absolut, galat positif dan galat negatif tidak saling menghilangkan. Dalam evaluasi forecasting,

MAE termasuk ukuran galat yang umum digunakan untuk menilai besar kesalahan prediksi (Hewamalage et al., 2023).

MAE dapat dibaca sebagai rata-rata besar kesalahan prediksi. Berbeda dari RMSE, MAE tidak menguadratkan galat. Oleh karena itu, MAE tidak memberi pengaruh sebesar RMSE terhadap galat yang sangat besar. Karakteristik ini membuat MAE berguna untuk membaca besar galat secara lebih langsung, karena nilainya menunjukkan rata-rata selisih absolut antara nilai aktual dan nilai prediksi.

Dalam penelitian ini, MAE digunakan sebagai metrik pendamping RMSE. RMSE tetap menjadi metrik utama karena lebih sensitif terhadap galat besar, sedangkan MAE membantu membaca rata-rata besar kesalahan prediksi secara lebih sederhana. MAE digunakan sebagai pelengkap karena dapat menunjukkan rata-rata besar galat absolut secara lebih langsung. Penggunaannya bersama RMSE juga ditemukan dalam penelitian prediksi saham dan *equity premium*, sehingga MAE sesuai digunakan sebagai metrik pendamping dalam evaluasi model prediksi *return* (Cao, 2025).

2.7.3 Alasan Tidak Menggunakan MAPE dan R-squared

Penelitian ini tidak menggunakan MAPE dan R-squared sebagai metrik evaluasi utama karena keduanya memiliki keterbatasan untuk konteks target *return* IHSG. MAPE atau *Mean Absolute Percentage Error* adalah metrik berbasis persentase yang menghitung kesalahan prediksi terhadap nilai aktual. Karena perhitungannya membagi galat dengan nilai aktual, MAPE dapat bermasalah ketika nilai aktual sama dengan nol atau sangat dekat dengan nol. Pada kondisi tersebut, nilai MAPE dapat menjadi tidak terdefinisi atau membesar secara tidak proporsional (Hewamalage et al., 2023; Kim & Kim, 2016).

Keterbatasan tersebut relevan dengan penelitian ini karena target yang diprediksi adalah *return*. Nilai *return* dapat berukuran kecil, mendekati nol, atau berubah tanda dari positif menjadi negatif. Jika MAPE digunakan pada kondisi seperti itu, nilai kesalahan persentase dapat menjadi tidak stabil dan sulit ditafsirkan. Oleh karena itu, MAPE tidak digunakan sebagai metrik utama dalam penelitian ini. Keputusan ini bukan berarti MAPE selalu tidak tepat untuk semua penelitian, tetapi MAPE kurang sesuai untuk target *return* yang dapat bernilai kecil atau mendekati nol.

R-squared juga tidak digunakan sebagai metrik utama. R-squared umumnya digunakan untuk melihat seberapa besar variasi nilai aktual dapat dijelaskan oleh model. Namun, pada data *time series* dan prediksi *return* saham *out-of-sample*, R-squared dapat memberi kesan

evaluasi yang kurang tepat. Pada konteks prediksi *return* saham, *out-of-sample* R-squared dapat bernilai rendah atau negatif ketika model tidak lebih baik daripada metode pembanding sederhana seperti *historical average benchmark* (Campbell & Thompson, 2008; Hewamalage et al., 2023).

Fokus penelitian ini adalah mengukur besar galat prediksi *return*, bukan menjelaskan variasi data secara umum. Karena itu, RMSE dan MAE lebih sesuai digunakan sebagai metrik evaluasi. RMSE digunakan sebagai metrik utama karena lebih sensitif terhadap galat besar, sedangkan MAE digunakan sebagai metrik pendamping untuk membaca rata-rata besar galat absolut. R-squared tetap dapat berguna pada penelitian lain, tetapi dalam penelitian ini tidak dijadikan metrik utama karena fokus evaluasinya adalah besar kesalahan prediksi *return*. Penggunaan RMSE dan MAE juga sejalan dengan penelitian prediksi saham yang memakai metrik galat untuk mengevaluasi performa model (Cao, 2025).

Nilai RMSE dan MAE hanya menunjukkan besar galat rata-rata, tetapi tidak menunjukkan apakah selisih galat antarmodel cukup besar untuk disebut berarti. Ketika selisihnya kecil, perbedaan tersebut dapat berasal dari variasi acak pada sampel pengujian yang terbatas. Karena itu, panduan evaluasi peramalan menempatkan pengujian signifikansi sebagai salah satu langkah evaluasi setelah galat dihitung (Hewamalage et al., 2023).

Uji Diebold-Mariano dirancang untuk membandingkan akurasi dua prediksi pada data dan periode yang sama. Uji ini bekerja pada selisih nilai *loss* kedua prediksi di setiap titik pengamatan, lalu menilai apakah rata-rata selisih tersebut berbeda dari nol. Karena yang dibandingkan adalah prediksi dan bukan model secara umum, uji ini sesuai untuk membandingkan hasil dua metode pada rangkaian pengujian yang identik (Diebold & Mariano, 2002; Hewamalage et al., 2023).

Pada prediksi dengan horizon lebih dari satu langkah, pengamatan yang berdekatan dapat memuat periode target yang saling tumpang tindih, sehingga selisih *loss* antarpengamatan berkorelasi. Kondisi tersebut membuat penaksir ragam sederhana menjadi kurang tepat, sehingga digunakan penaksir ragam yang tahan terhadap korelasi antarpengamatan dan ragam yang tidak konstan (Newey & West, 1987).

Persoalan berikutnya muncul ketika banyak pasangan diuji sekaligus. Semakin banyak pengujian dilakukan, semakin besar peluang menemukan hasil yang tampak signifikan padahal hanya kebetulan. Koreksi perbandingan ganda digunakan untuk menahan peluang tersebut, salah satunya melalui prosedur bertahap yang mengurutkan nilai p lalu menyesuaikan ambang keputusannya (Holm, 1979). Penelitian prediksi saham menunjukkan bahwa bukti performa

yang semula tampak meyakinkan dapat gugur setelah koreksi semacam itu diterapkan (Pagliaro, 2026).

Uji Diebold-Mariano sudah digunakan pada penelitian peramalan deret keuangan maupun pada prediksi *return* saham Indonesia untuk menilai apakah selisih akurasi antarmetode signifikan (Beck et al., 2025; Indra et al., 2025).

Dalam penelitian ini, uji tersebut digunakan untuk menilai perbedaan akurasi antarmodel dan antardataset. Fungsi *loss* yang dipakai, panjang koreksi ragam, serta pengelompokan koreksi perbandingan ganda dijelaskan pada Bab 3.

2.8 Penelitian Terdahulu

Penelitian terdahulu yang berkaitan dengan penelitian ini dapat dikelompokkan ke dalam beberapa arah. Pertama, terdapat penelitian yang membahas prediksi IHSG atau JCI menggunakan metode sederhana, model statistik, dan model *machine learning*. Kedua, terdapat penelitian yang membahas prediksi *return* saham Indonesia menggunakan indikator teknikal. Ketiga, terdapat penelitian yang membahas volatilitas dan penggunaan model ARCH atau GARCH. Keempat, terdapat penelitian yang lebih menekankan aspek metodologis, seperti penggunaan LightGBM, *walk-forward validation*, metode pembanding sederhana, dan evaluasi berbasis rezim pasar. Pengelompokan ini digunakan agar penelitian terdahulu dapat dibaca berdasarkan hubungannya dengan penelitian ini, bukan hanya berdasarkan kesamaan topik pasar saham.

Ringkasan penelitian terdahulu yang digunakan dalam penelitian ini ditunjukkan pada Tabel 2.1.

Tabel 2.1 Penelitian Terdahulu

Penulis	Judul Penelitian	Model	Fokus	Hasil
(Ismanto, 2026)	Forecasting the Jakarta Composite Index (IHSG) Using the Moving Average Method	Simple Moving Average	Prediksi/peramalan IHSG berbasis data historis indeks	Moving average dapat menangkap tren umum IHSG, tetapi hasil peramalan dipengaruhi oleh ukuran jendela yang digunakan.

Penulis	Judul Penelitian	Model	Fokus	Hasil
(Najibullah et al., 2024)	Prediksi Indeks Harga Saham Gabungan (IHSG) Menggunakan Regresi K-Nearest Neighbors	K-Nearest Neighbors Regression	Prediksi nilai IHSG menggunakan variabel eksternal seperti harga emas, kurs, dan indeks global	KNN Regression digunakan untuk memprediksi nilai IHSG dan menunjukkan bahwa variabel eksternal dapat dimanfaatkan sebagai masukan model. Target penelitian ini berupa level indeks, bukan *return*.
(Galena et al., 2025)	Prediction Analysis of Jakarta Composite Index Movement Using Support Vector Regression Method	Support Vector Regression	Prediksi pergerakan JCI/IHSG mingguan berbasis data historis	SVR digunakan untuk memprediksi nilai JCI/IHSG. Paper ini relevan sebagai contoh penggunaan *machine learning* pada indeks saham, tetapi targetnya masih berupa level indeks.
(Adinda et al., 2026)	Forecasting the IHSG with an ARIMA-ARCH Model to Address Heteroscedasticity Effects	ARIMA-ARCH	Prediksi IHSG dengan memperhatikan heteroskedastisitas pada data deret waktu	ARIMA-ARCH digunakan untuk menangani pola heteroskedastisitas pada data IHSG. Paper ini relevan untuk menunjukkan bahwa karakteristik volatilitas perlu diperhatikan, tetapi targetnya bukan *return* 20 sesi.
(Wiratama et al., 2025)	GARCH Model Hybridization with Feed Forward Neural Network Algorithm Approach for Predicting the Volatility of the Composite Stock Price Index	GARCH/EGARCH dan Feed Forward Neural Network	Prediksi volatilitas IHSG	Model hybrid GARCH/EGARCH dan neural network digunakan untuk memprediksi volatilitas IHSG. Paper ini relevan untuk mendukung penggunaan informasi volatilitas, tetapi tidak disamakan dengan penelitian ini karena targetnya adalah volatilitas, bukan *return*.

Penulis	Judul Penelitian	Model	Fokus	Hasil
(Indra et al., 2025)	Stock Return Prediction on the LQ45 Market Index in the Indonesia Stock Exchange Using a Machine Learning Algorithm Based on Technical Indicators	Linear Regression, Ridge Regression, Random Forest, dan XGBoost	Prediksi *return* saham LQ45 berbasis indikator teknikal	Paper ini menunjukkan bahwa prediksi *return* saham Indonesia dapat diuji menggunakan fitur teknikal dan model *machine learning*. Paper ini paling dekat dengan penelitian ini karena sama-sama membahas *return*, fitur teknikal, model *machine learning*, dan pentingnya metode pembanding sederhana.
(Saputra et al., 2025)	Indonesian Banking Stock Portfolio Optimization Based on Ridge Regression Prediction	Ridge Regression	Prediksi *return* saham perbankan Indonesia dan optimasi portofolio	Prediksi *return* saham perbankan Indonesia dan optimasi portofolio
(Pagliaro, 2026)	Regime-Aware LightGBM for Stock Market Forecasting: A Validated Walk-Forward Framework with Statistical Rigor and Explainable AI Analysis	LightGBM dengan *walk-forward validation* dan evaluasi berbasis rezim	Evaluasi model berbasis rezim pada saham NASDAQ-100	Paper ini relevan secara metodologis karena menggunakan LightGBM, *walk-forward validation*, metode pembanding sederhana, dan evaluasi berbasis rezim. Namun, paper ini tidak dijadikan pembanding objek karena tidak meneliti IHSG dan targetnya bukan *return* kontinu IHSG.

Berdasarkan Tabel 2.1, sebagian penelitian terdahulu pada IHSG atau JCI masih berfokus pada prediksi level indeks. Penelitian terlampir menunjukkan bahwa IHSG/JCI telah digunakan sebagai objek peramalan dengan metode sederhana, model statistik, dan *machine learning*. Namun, target yang digunakan pada penelitian tersebut berbeda dari penelitian ini. Beberapa penelitian memprediksi nilai indeks atau pergerakan indeks, sedangkan penelitian ini memprediksi *return* IHSG 20 sesi perdagangan ke depan. Perbedaan target ini penting karena evaluasi model pada prediksi level indeks tidak dapat disamakan langsung dengan evaluasi prediksi *return*.

Penelitian yang lebih dekat dengan penelitian ini adalah penelitian prediksi *return* saham Indonesia. (Indra et al., 2025) menggunakan indikator teknikal dan beberapa model *machine learning*, seperti Ridge Regression, Random Forest, dan XGBoost, untuk memprediksi *return* saham LQ45. Paper tersebut relevan karena sama-sama membahas *return*, fitur teknikal, model *machine learning*, dan pentingnya metode pembanding sederhana. Meskipun demikian, objek penelitian tersebut adalah saham-saham LQ45, bukan IHSG. Oleh karena itu, paper tersebut digunakan sebagai rujukan utama untuk konteks prediksi *return* saham Indonesia, tetapi bukan sebagai pembanding langsung terhadap hasil penelitian ini.

Selain itu, (Saputra et al., 2025) menggunakan Ridge Regression pada saham perbankan Indonesia. Penelitian tersebut menunjukkan bahwa Ridge Regression dapat digunakan pada data saham Indonesia dengan fitur historis seperti *moving average* dan *rolling standard deviation*. Namun, fokus penelitian tersebut berbeda karena tidak hanya membahas prediksi *return*, tetapi juga optimasi portofolio. Dengan demikian, paper tersebut digunakan sebagai rujukan pendukung bahwa Ridge Regression pernah digunakan dalam konteks saham Indonesia, bukan sebagai patokan utama performa model.

Penelitian yang berhubungan dengan volatilitas juga menjadi bagian penting dalam kajian terdahulu. (Adinda et al., 2026) menggunakan ARIMA-ARCH untuk memperhatikan heteroskedastisitas pada data IHSG, sedangkan (Wiratama et al., 2025) menggunakan model berbasis GARCH/EGARCH dan neural network untuk memprediksi volatilitas IHSG. Kedua penelitian tersebut menunjukkan bahwa volatilitas merupakan karakteristik yang perlu diperhatikan pada data IHSG. Akan tetapi, arah penelitian tersebut berbeda dari penelitian ini.

Penelitian ini tetap memprediksi *return*, sedangkan informasi volatilitas digunakan sebagai fitur tambahan dan sebagai dasar untuk membaca perubahan galat pada rezim volatilitas.

Dari sisi metodologi, (Pagliaro, 2026) menjadi rujukan penting untuk penggunaan LightGBM, *walk-forward validation*, metode pembanding sederhana, dan evaluasi berbasis rezim pasar. Paper tersebut tidak digunakan sebagai pembanding objek karena meneliti saham NASDAQ-100, bukan IHSG. Namun, pendekatan metodologisnya relevan karena penelitian ini juga membutuhkan evaluasi yang mempertahankan urutan waktu dan membaca perubahan performa model pada kondisi pasar yang berbeda.

Berdasarkan penelitian terdahulu tersebut, masih terdapat ruang untuk mengevaluasi prediksi *return* IHSG dengan rancangan yang lebih terarah pada stabilitas performa model. Beberapa penelitian terdahulu masih memprediksi level indeks, beberapa menggunakan target

volatilitas, dan beberapa belum menekankan perbandingan model terhadap metode pembanding sederhana serta perubahan performa pada rezim volatilitas berbeda. Oleh karena itu, penelitian ini memosisikan diri pada evaluasi stabilitas performa empat model *machine learning* dalam prediksi *return* IHSG 20 sesi perdagangan ke depan. Evaluasi dilakukan dengan membandingkan beberapa rancangan dataset, metode pembanding sederhana, performa tahunan, dan performa pada rezim volatilitas rendah, sedang, dan tinggi.

2.9 Posisi Penelitian

Berdasarkan penelitian terdahulu yang telah dibahas, penelitian ini tidak ditempatkan sebagai upaya untuk menyatakan bahwa satu pendekatan pasti lebih baik daripada penelitian sebelumnya. Penelitian ini memiliki posisi yang berbeda karena berfokus pada evaluasi stabilitas performa model *machine learning* dalam prediksi *return* IHSG 20 sesi perdagangan ke depan. Dengan demikian, fokus penelitian tidak hanya berada pada pencarian model dengan galat rata-rata paling rendah, tetapi juga pada cara performa model berubah pada beberapa kondisi pengujian.

Perbedaan pertama terletak pada target prediksi. Beberapa penelitian terdahulu pada IHSG atau JCI masih berfokus pada prediksi level indeks atau nilai harga. Penelitian ini menggunakan *return* sebagai target prediksi, sehingga hasil prediksi dibaca sebagai perubahan relatif IHSG dalam horizon 20 sesi perdagangan. Pemilihan target *return* membuat penelitian ini tidak diarahkan untuk memprediksi nilai indeks secara langsung, tetapi untuk mengukur seberapa dekat prediksi model terhadap perubahan IHSG pada periode yang ditentukan.

Perbedaan kedua terletak pada model yang dibandingkan. Penelitian ini membandingkan Ridge Regression, Random Forest, LightGBM, dan XGBoost sebagai kandidat model prediksi *return* IHSG. Performa keempat model tersebut dinilai berdasarkan hasil evaluasi, bukan berdasarkan asumsi awal bahwa salah satu model pasti lebih unggul.

Perbedaan ketiga terletak pada rancangan dataset. Penelitian ini membandingkan tiga dataset untuk melihat apakah tambahan informasi makroekonomi dan volatilitas dapat membantu evaluasi prediksi *return* IHSG. Detail isi setiap dataset dijelaskan pada Bab 3.

Penelitian ini juga menempatkan metode pembanding sederhana sebagai bagian penting dari evaluasi. Model *machine learning* tidak hanya dibandingkan dengan sesama model *machine learning*, tetapi juga dibandingkan dengan *no-change forecast* dan *persistence forecast*. Perbandingan ini diperlukan agar performa model dapat dibaca secara lebih hati-hati. Jika model hanya lebih baik daripada model lain, tetapi tidak lebih baik daripada aturan

prediksi sederhana, maka manfaat tambahan dari model perlu ditafsirkan dengan lebih terbatas. Selain itu, penelitian ini menggunakan evaluasi berbasis urutan waktu. Evaluasi tidak dilakukan dengan mencampur data masa lalu dan masa depan, tetapi mempertahankan karakter data *time series*. *Chronological holdout* digunakan sebagai evaluasi awal, sedangkan *walk-forward validation* digunakan sebagai evaluasi utama. Dengan cara ini, performa model dapat dibaca pada beberapa periode uji, bukan hanya pada satu pembagian data.

Selain itu, penelitian ini menempatkan pengujian statistik sebagai bagian dari evaluasi. Selisih galat antarmodel tidak hanya dilaporkan sebagai angka, tetapi diuji untuk menilai apakah selisih tersebut cukup kuat untuk disebut berarti, disertai koreksi ketika banyak perbandingan diuji sekaligus. Kontribusi setiap kelompok fitur juga dinilai secara terpisah melalui *ablation study*, sehingga perubahan hasil dapat dikaitkan dengan kelompok fitur yang ditambahkan. Ketiga dataset dievaluasi pada periode, tanggal, dan pembagian data yang sama agar perbedaan hasil tidak berasal dari perbedaan cakupan periode. Ketiga hal tersebut membedakan penelitian ini dari sebagian penelitian terdahulu yang melaporkan selisih galat tanpa pengujian, menilai kelompok fitur secara gabungan, atau membandingkan dataset pada periode yang tidak sama.

Posisi utama penelitian ini terletak pada evaluasi stabilitas performa. Stabilitas performa dalam penelitian ini tidak dimaknai sebagai jaminan bahwa model selalu baik pada semua kondisi. Stabilitas performa dibaca sebagai kemampuan model mempertahankan galat yang relatif rendah ketika diuji pada periode waktu dan rezim volatilitas yang berbeda. Oleh karena itu, penelitian ini mengevaluasi performa model secara tahunan dan berdasarkan rezim volatilitas rendah, sedang, dan tinggi.

Volatilitas dalam penelitian ini tidak digunakan sebagai target prediksi utama. Target penelitian tetap berupa *return* IHSG 20 sesi perdagangan ke depan. Volatilitas digunakan untuk mengelompokkan kondisi pasar dan membaca apakah galat prediksi berubah ketika pasar berada pada kondisi yang lebih tenang atau lebih bergejolak. Dengan demikian, analisis rezim volatilitas dalam penelitian ini bersifat evaluatif dan deskriptif, bukan untuk menyimpulkan hubungan kausal.

Berdasarkan posisi tersebut, penelitian ini berada pada ruang evaluasi model, bukan pada pembuktian profitabilitas investasi. Penelitian ini tidak menguji strategi *trading* atau menyatakan bahwa model yang diuji dapat langsung digunakan untuk keputusan investasi.

Fokus penelitian adalah membandingkan performa model, rancangan dataset, metode pembanding sederhana, performa tahunan, dan perubahan galat pada rezim volatilitas berbeda.

Rancangan penelitian yang digunakan untuk menjalankan posisi tersebut dijelaskan pada Bab 3.

BAB III
METODOLOGI PENELITIAN

3.1 Desain Penelitian

Penelitian ini merupakan penelitian kuantitatif dengan desain evaluasi komparatif *out-of-sample*. Pendekatan kuantitatif digunakan karena performa model dinilai berdasarkan hasil pengukuran numerik, sedangkan desain komparatif digunakan untuk membandingkan beberapa model dan susunan dataset dalam kondisi pengujian yang sama. Evaluasi *out-of-sample* berarti model diuji menggunakan data yang tidak dilibatkan dalam proses pelatihan maupun pemilihan *hyperparameter*. Pendekatan tersebut dipilih agar proses evaluasi lebih mendekati kondisi prediksi yang sebenarnya, yaitu ketika model hanya menggunakan informasi dari masa lalu untuk memprediksi nilai yang belum diketahui. Pengujian pada data baru juga diperlukan karena performa yang baik pada data pelatihan belum tentu dapat dipertahankan ketika model digunakan pada periode yang berbeda (Tashman, 2000).

Fokus penelitian ini adalah memprediksi *return* Indeks Harga Saham Gabungan (IHSG) untuk 20 sesi perdagangan ke depan dengan membandingkan Ridge Regression, Random Forest, XGBoost, dan LightGBM. Penelitian ini juga menggunakan tiga susunan dataset untuk melihat perubahan performa setelah informasi baru ditambahkan. Dataset A berisi fitur teknikal IHSG, Dataset B menambahkan indikator ekonomi domestik, sedangkan Dataset C menambahkan indikator pasar global dan volatilitas hasil estimasi GARCH. Rincian variabel, sumber data, dan pembentukan dataset dijelaskan pada Subbab 3.2 dan 3.3.

Evaluasi utama dilakukan pada periode uji bersama agar seluruh model dan dataset dinilai pada kondisi pasar serta tanggal prediksi yang sama. Periode pengujian terdiri dari enam *fold* tahunan, yaitu tahun 2020 sampai dengan 2025. Prediksi dilakukan satu kali pada setiap bulan kalender untuk mengurangi tumpang tindih target *return* 20 sesi perdagangan. Rincian tanggal evaluasi dan pembagian *fold* dijelaskan pada Subbab 3.4.

Pembagian data dilakukan menggunakan *expanding walk-forward validation* dengan tetap mempertahankan urutan waktu. Pada setiap *fold*, model menggunakan data sebelum tahun pengujian dan periode pelatihan bertambah pada *fold* berikutnya. Proses pemisahan data juga mempertimbangkan *target date* agar target 20 sesi pada data pelatihan tidak mencapai periode validasi atau pengujian. Prosedur pembagian data dan *purging* dijelaskan pada Subbab 3.4, sedangkan penalaan model dijelaskan pada Subbab 3.5.

Kondisi volatilitas dikelompokkan menjadi rezim rendah, sedang, dan tinggi menggunakan *Markov Switching*. Pengelompokan tidak menggunakan batas kuantil yang tetap, tetapi berdasarkan varians dan probabilitas *state* yang diestimasi dari data yang telah tersedia. Prosedur pembentukan rezim dan perlakuan terhadap jumlah observasi yang terbatas dijelaskan pada Subbab 3.7.

Performa utama model diukur menggunakan *Root Mean Squared Error* (RMSE), sedangkan *Mean Absolute Error* (MAE) dan *Directional Accuracy* (DA) digunakan sebagai ukuran pendukung. Evaluasi juga mencakup stabilitas antar-*fold* dan rezim, perbandingan dengan metode sederhana, *ablation study*, serta uji Diebold-Mariano. Tata cara perhitungan dan analisis tersebut dijelaskan pada Subbab 3.6 sampai dengan Subbab 3.9.

Berdasarkan rancangan tersebut, hasil penelitian digunakan untuk menjelaskan perbedaan akurasi dan stabilitas performa model pada periode serta rezim volatilitas yang berbeda. Hasil penambahan fitur digunakan untuk menilai perubahan kemampuan prediksi, tetapi tidak digunakan sebagai bukti bahwa suatu fitur menyebabkan perubahan *return* IHSG. Uji statistik yang digunakan juga hanya menunjukkan kekuatan perbedaan performa prediksi dan tidak membuktikan profitabilitas model dalam strategi investasi. Oleh karena itu, penelitian ini berada dalam ruang lingkup prediktif dan deskriptif serta tidak ditujukan untuk menarik kesimpulan sebab-akibat. Alur keseluruhan penelitian ditunjukkan pada Gambar 3.1.



Gambar 3.1 Alur Eksperimental Penelitian

3.2 Sumber dan Pengambilan Data

Data perdagangan harian IHSG diperoleh dari Yahoo Finance melalui pustaka Python *yfinance* menggunakan simbol ^JKSE. Data mentah yang diambil mencakup tanggal perdagangan, harga pembukaan (*Open*), harga tertinggi (*High*), harga terendah (*Low*), harga penutupan (*Close*), harga penutupan disesuaikan (*Adjusted Close*), dan volume perdagangan (*Volume*). Kolom *Adjusted Close* tetap disimpan sebagai bagian dari data mentah, sedangkan data harga dan volume lainnya digunakan dalam proses pembentukan fitur dan target. Penjelasan mengenai penggunaan setiap kolom dalam pembentukan fitur dan target disampaikan pada Subbab 3.3.

Data tambahan yang digunakan terdiri dari indikator ekonomi domestik dan indikator pasar global. Inflasi dan suku bunga Indonesia diperoleh dari Federal Reserve Economic Data (FRED). Nilai tukar USD/IDR, S&P 500, *Dollar Index*, dan *US 10-Year Treasury Yield*

diperoleh dari Yahoo Finance melalui *yfinance*. Ringkasan variabel, sumber, frekuensi, aturan waktu, dan penggunaannya pada setiap dataset ditunjukkan pada Tabel 3.1.

Pengambilan data dilakukan secara otomatis menggunakan skrip Python, kemudian data mentah disimpan sebagai *snapshot* penelitian. Data eksternal disesuaikan dengan tanggal perdagangan IHSG sebelum digunakan dalam pembentukan dataset. Untuk data harian, nilai terakhir yang telah tersedia diteruskan sampai terdapat observasi baru, kemudian variabel diberi *lag* satu sesi perdagangan. Data inflasi dan suku bunga yang memiliki frekuensi bulanan terlebih dahulu digeser satu bulan kalender sebagai pendekatan konservatif terhadap waktu ketersediaan data. Setelah diselaraskan dengan kalender perdagangan IHSG, kedua variabel tersebut juga diberi *lag* satu sesi perdagangan. Data mentah IHSG beserta seluruh variabel eksternal tersedia sejak Mei 2011, sedangkan awal periode dataset penelitian ditentukan oleh kebutuhan data awal estimasi GARCH dan penyelarasan tanggal antardataset yang dijelaskan pada Subbab 3.3.

Tabel 3.1 Ringkasan Sumber Data

Variabel	Sumber	Frekuensi	Aturan waktu/*lag*	Dataset
IHSG	Yahoo Finance melalui *yfinance*	Harian perdagangan	Nilai pada tanggal perdagangan	A, B, C
Nilai tukar USD/IDR	Yahoo Finance melalui *yfinance*	Harian	*Lag* satu sesi perdagangan	B, C
Inflasi Indonesia	FRED	Bulanan	Digeser satu bulan dan diberi *lag* satu sesi	B, C
Suku bunga Indonesia	FRED	Bulanan	Digeser satu bulan dan diberi *lag* satu sesi	B, C
S&P 500	Yahoo Finance melalui *yfinance*	Harian	*Lag* satu sesi perdagangan	C
Dollar Index	Yahoo Finance melalui *yfinance*	Harian	*Lag* satu sesi perdagangan	C
US 10-Year Treasury Yield	Yahoo Finance melalui *yfinance*	Harian	*Lag* satu sesi perdagangan	C
Volatilitas GARCH	Diolah dari *return* IHSG	Harian	*Lag* satu sesi perdagangan	C

Volatilitas GARCH pada Tabel 3.1 tidak diperoleh dari sumber data eksternal, tetapi dihasilkan dari proses estimasi menggunakan data *return* IHSG. Variabel tersebut dicantumkan dalam tabel untuk menunjukkan asal data dan penggunaannya pada Dataset C. Prosedur pembentukan dan pemeriksaan hasil estimasi GARCH dijelaskan pada Subbab 3.3.

3.3 Pembentukan Dataset, Fitur, dan Target

Pembentukan dataset diawali dengan mengurutkan data IHSG berdasarkan tanggal perdagangan, menghapus duplikasi tanggal, mengubah kolom harga dan volume menjadi numerik, serta menghapus baris dengan nilai *Open*, *High*, *Low*, *Close*, atau *Volume* yang tidak valid. Tahap ini dilakukan sebelum pembentukan fitur agar setiap observasi memiliki urutan waktu dan nilai dasar yang konsisten.

Target penelitian adalah *return* IHSG 20 sesi perdagangan ke depan. Nilai tersebut dihitung dari harga penutupan pada sesi saat prediksi dibuat dan harga penutupan 20 sesi setelahnya dengan rumus berikut.

$$R_{t,20} = \frac{\text{Close}_{t+20}}{\text{Close}_t} - 1 \tag{3.1}$$

$R_{t,20}$ menunjukkan *return* IHSG untuk 20 sesi berikutnya, Close_t merupakan harga penutupan pada sesi t, dan Close_{t+20} merupakan harga penutupan pada sesi t+20. Tanggal t+20 disimpan sebagai *target date* untuk mencegah penggunaan target yang belum selesai pada proses validasi. Target berbentuk nilai kontinu sehingga pemodelan dilakukan sebagai masalah regresi.

Fitur teknikal dibentuk dari informasi harga dan volume yang telah tersedia sebelum sesi prediksi. Sebanyak 59 fitur dibagi ke dalam enam kelompok sebagaimana ditunjukkan pada Tabel 3.2.

Tabel 3.2 Kelompok Fitur Teknikal

Kelompok Fitur	Fitur dan Keterangan	Jumlah
Harga dan *return* dasar	*Return* satu sesi, log harga, log volume, perubahan volume, rentang *High-Low*, dan *return Open-Close*	6
Lag historis	Harga penutupan, *return* satu sesi, dan perubahan volume pada *lag* 1, 2, 3, 5, 10, dan 20 sesi	18
Statistik *rolling* dan momentum	Rata-rata harga, rata-rata *return*, dan volatilitas pada jendela 5, 10, 20, dan 60 sesi, serta momentum 5, 10, dan 20 sesi	15
Tren	EMA (*Exponential Moving Average*) memberi bobot lebih besar pada data terbaru. MACD (*Moving Average Convergence Divergence*) mengukur selisih dua EMA dan dilengkapi *Bollinger Bands*	11
Volatilitas dan osilator	ATR (*Average True Range*) mengukur rentang pergerakan. RSI (*Relative Strength Index*) dan *stochastic oscillator* mengukur kekuatan relatif pergerakan	5
Volume	Rata-rata dan rasio volume, serta OBV (*On-Balance Volume*) yang mengakumulasi volume berdasarkan arah perubahan harga	4
Total		59

Lag 1, 2, 3, 5, 10, dan 20 sesi digunakan sebagai masukan historis pada waktu t, sedangkan 20 sesi pada target menunjukkan periode masa depan yang diprediksi. Keduanya memiliki fungsi berbeda. Variasi *lag* membantu model menangkap pola jangka sangat pendek hingga sekitar satu bulan perdagangan tanpa menggunakan informasi setelah waktu prediksi.

Dataset C juga menggunakan volatilitas hasil estimasi GARCH(1,1) dengan rata-rata konstan dan distribusi *Student-t*. Estimasi dilakukan secara *expanding* menggunakan *return* IHSG yang telah tersedia, dengan minimum 504 observasi dan pembaruan parameter setiap 21 sesi. Volatilitas satu langkah ke depan dihitung berdasarkan persamaan berikut menurut (Engle, 1982) dan (Bollerslev, 1986). Batas minimum 504 observasi tersebut setara dengan sekitar dua tahun sesi perdagangan.

$$\varepsilon_t = \sigma_t z_t, \sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{3.2}$$

Keterangan:
ε_t = galat *return* IHSG pada waktu t.
σ_t = simpangan baku bersyarat, yaitu volatilitas bersyarat, pada waktu t.
z_t = variabel acak baku berdistribusi *Student-t*.
ω = konstanta varians bersyarat.
α = koefisien ARCH, yaitu bobot galat kuadrat satu periode sebelumnya.
β = koefisien GARCH, yaitu bobot varians bersyarat satu periode sebelumnya.

Hasil estimasi dinyatakan valid apabila proses konvergen, seluruh parameter bernilai terbatas, $\omega > 0$, α dan β tidak negatif, serta $\alpha + \beta < 1$. Jika estimasi dengan distribusi *Student-t* tidak valid, model diestimasi kembali menggunakan distribusi normal. Parameter valid terakhir digunakan apabila estimasi ulang tetap gagal. Diagnostik residual dilakukan pada *lag* 10 menggunakan uji Ljung-Box pada residual terstandarisasi dan kuadratnya, ARCH-LM, serta Jarque-Bera. Status konvergensi, parameter, dan hasil diagnostik disimpan oleh *pipeline* untuk dilaporkan pada Bab IV.

Fitur teknikal kemudian digabungkan dengan variabel eksternal pada Tabel 3.1 sesuai tanggal dan aturan waktu yang telah ditetapkan. Baris yang belum memiliki fitur atau target lengkap dihapus, kemudian ketiga dataset diselaraskan pada irisan tanggal yang sama sehingga masing-masing memiliki 3.023 observasi. Ketiga dataset menggunakan target dan periode observasi yang sama agar perbandingan performa tidak dipengaruhi oleh perbedaan cakupan data. Komposisi dataset ditunjukkan pada Tabel 3.3.

Tabel 3.3 Ringkasan Dataset Penelitian

Dataset	Komposisi	Jumlah Fitur	Observasi
Dataset A	59 fitur teknikal	59	3.023
Dataset B	59 fitur teknikal dan 3 indikator makroekonomi domestik	62	3.023
Dataset C	59 fitur teknikal, 3 indikator domestik, 3 indikator pasar global, dan 1 fitur GARCH	66	3.023

3.4 Pembagian Data dan Prapemrosesan

Setelah Dataset A, Dataset B, dan Dataset C diselaraskan, data evaluasi dibentuk dengan frekuensi bulanan. Pada setiap bulan kalender, *pipeline* memilih observasi terakhir yang tersedia dan telah memiliki fitur serta target *return* 20 sesi secara lengkap. Tanggal tersebut digunakan sebagai *monthly anchor*. Pemilihan ini hanya mengubah frekuensi observasi yang digunakan untuk pelatihan dan pengujian, sedangkan fitur teknikal tetap dibentuk dari data historis harian.

Pemilihan satu *anchor* setiap bulan memberi jarak yang lebih lebar antarprediksi sehingga potensi tumpang tindih target dapat dikurangi. Jarak tersebut tidak selalu menghilangkan seluruh tumpang tindih karena jumlah sesi perdagangan pada setiap bulan dapat berbeda. Evaluasi utama menggunakan 72 observasi dari Januari 2020 sampai Desember 2025, dengan 12 observasi pada setiap tahun untuk masing-masing dataset. Ringkasan protokol evaluasi utama ditunjukkan pada Tabel 3.4.

Tabel 3.4 Ringkasan Protokol Evaluasi Utama

Komponen	Ketentuan
Dataset	Dataset A, Dataset B, dan Dataset C
Frekuensi evaluasi	Satu observasi pada setiap bulan kalender
Pemilihan *anchor*	Observasi terakhir yang tersedia dengan fitur dan target lengkap
Target	*Return* IHSG 20 sesi perdagangan ke depan
Awal data pelatihan	28 Juni 2013
Periode pengujian	Januari 2020 sampai Desember 2025
Jumlah *fold*	Enam *fold* tahunan
Observasi pengujian	72 observasi untuk setiap dataset

Pembagian data dilakukan menggunakan *expanding walk-forward validation*. Pada setiap *fold*, data pelatihan terdiri dari seluruh *monthly anchor* sebelum tahun pengujian,

sedangkan data pada tahun pengujian digunakan sebagai data uji. Periode pelatihan bertambah ketika evaluasi berpindah ke *fold* berikutnya. Rasio tetap seperti 50:50 atau 70:30 tidak digunakan karena penelitian memerlukan enam periode pengujian tahunan dengan tanggal yang sama pada ketiga dataset.

Periode pada Tabel 3.5 menunjukkan *monthly anchor* pertama dan terakhir pada setiap bagian data. Periode pelatihan yang dicantumkan merupakan periode setelah proses *purging*, sedangkan periode uji menunjukkan *anchor* pertama dan terakhir pada tahun pengujian.

Tabel 3.5 Pembagian Data Expanding Walk-Forward Validation

Fold	Tahun Uji	Periode Data Pelatihan	Periode Data Uji	Observasi Uji
1	2020	28 Juni 2013 sampai 17 Desember 2019	27 Januari sampai 30 Desember 2020	12
2	2021	28 Juni 2013 sampai 30 November 2020	29 Januari sampai 30 Desember 2021	12
3	2022	28 Juni 2013 sampai 30 Desember 2021	31 Januari sampai 30 Desember 2022	12
4	2023	28 Juni 2013 sampai 30 Desember 2022	31 Januari sampai 29 Desember 2023	12
5	2024	28 Juni 2013 sampai 29 Desember 2023	31 Januari sampai 30 Desember 2024	12
6	2025	28 Juni 2013 sampai 29 November 2024	31 Januari sampai 30 Desember 2025	12

Proses *purging* dilakukan berdasarkan tanggal berakhirnya target pada setiap observasi. Suatu observasi hanya dipertahankan dalam data pelatihan apabila tanggal target 20 sesinya lebih awal daripada tanggal pertama periode uji. Observasi yang targetnya mencapai periode pengujian dikeluarkan agar nilai target pelatihan tidak menggunakan harga dari periode uji. Jumlah observasi yang dikeluarkan tidak ditetapkan sebanyak 20 baris karena data evaluasi menggunakan frekuensi bulanan, sedangkan horizon target dihitung dalam sesi perdagangan.

Pemilihan konfigurasi model menggunakan validasi internal pada data pelatihan setiap *fold*. Sebanyak 80% observasi awal digunakan sebagai bagian *fit*, sedangkan 20% observasi terakhir digunakan sebagai bagian validasi. Pembagian dilakukan secara kronologis dan bagian *fit* kembali dipurge berdasarkan tanggal target terhadap awal periode validasi. Dengan demikian, data uji tahunan tidak digunakan dalam pemilihan konfigurasi model.

Prapemrosesan dilakukan secara terpisah pada setiap *fold* setelah data pelatihan dan data uji dibentuk. Tahap pertama adalah *winsorization* pada fitur masukan menggunakan batas

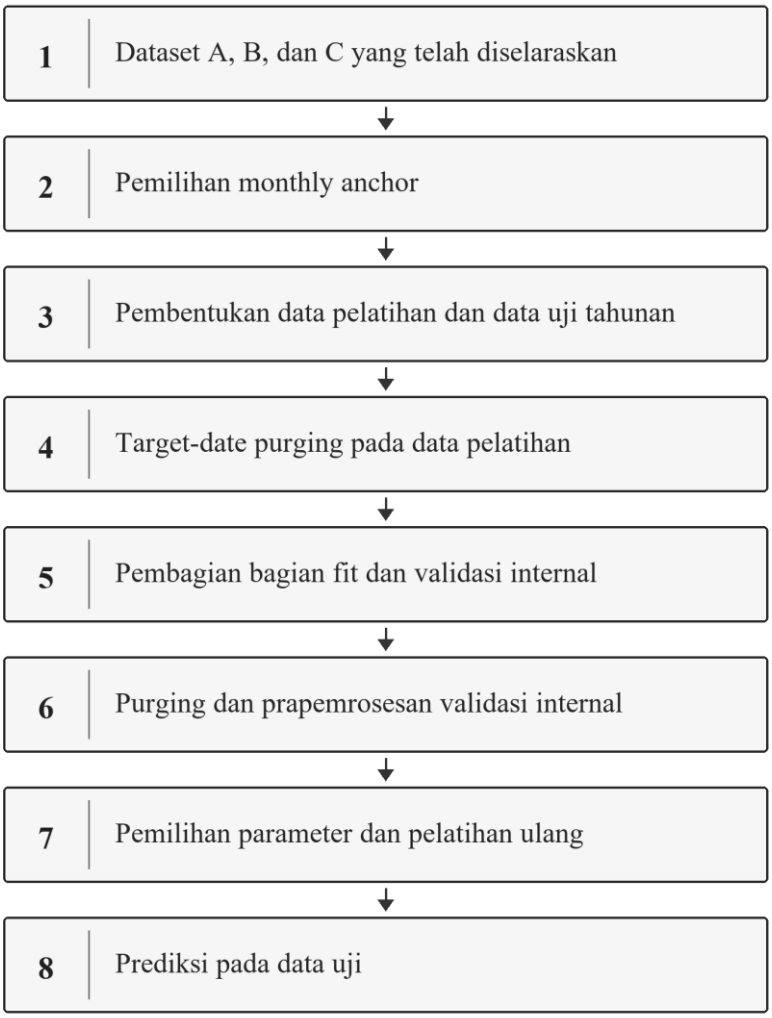
persentil 1% dan 99% yang dihitung hanya dari data pelatihan. Batas tersebut kemudian diterapkan pada data pelatihan, validasi, dan pengujian. Target *return* tidak dikenai *winsorization* agar evaluasi tetap menggunakan nilai aktualnya.

Standardisasi diterapkan khusus pada Ridge Regression melalui StandardScaler yang menjadi bagian dari *pipeline* model. Standardisasi diperlukan karena Ridge Regression memberikan penalti terhadap koefisien, sedangkan fitur penelitian memiliki skala yang berbeda. Nilai fitur yang telah distandardisasi dihitung menggunakan rumus berikut.

$$z_{i,j} = \frac{x_{i,j} - \mu_{train,j}}{\sigma_{train,j}} \tag{3.3}$$

Keterangan:
$z_{i,j}$ = nilai fitur ke-j pada observasi ke-i setelah standardisasi.
$x_{i,j}$ = nilai fitur asli.
$\mu_{train,j}$ = rata-rata fitur ke-j pada data pelatihan.
$\sigma_{train,j}$ = simpangan baku fitur ke-j pada data pelatihan.

Rata-rata dan simpangan baku dihitung hanya dari data pelatihan, kemudian nilai yang sama digunakan untuk mentransformasi data validasi dan data uji. Random Forest, XGBoost, dan LightGBM tidak menggunakan standardisasi karena ketiga model tersebut membentuk prediksi berdasarkan pemisahan nilai fitur pada struktur pohon. Urutan pembagian data dan prapemrosesan pada setiap *fold* ditunjukkan pada Gambar 3.2.



Gambar 3.2 Alur Pembagian Data dan Prapemrosesan pada Setiap Fold

3.5 Model dan Metode Pembanding

Penelitian ini membandingkan empat model regresi, yaitu Ridge Regression, Random Forest, LightGBM, dan XGBoost. Pemilihan model tersebut dilakukan untuk membandingkan model linear dengan beberapa model berbasis pohon yang mampu menangkap hubungan nonlinier. Keempat model diperlakukan sebagai kandidat yang setara sehingga penentuan model terbaik didasarkan pada hasil pengujian *out-of-sample*.

Ridge Regression merupakan model regresi linear yang menggunakan regularisasi L2. Regularisasi membatasi besarnya koefisien model sehingga pengaruh fitur yang saling berkorelasi dapat dikendalikan. Karakteristik tersebut membuat Ridge Regression digunakan sebagai model linear yang relatif sederhana dalam penelitian ini. Penggunaan Ridge Regression pada prediksi *return* saham Indonesia juga ditemukan dalam penelitian (Indra et al., 2025).

Random Forest merupakan model *tree ensemble* yang menggabungkan hasil prediksi dari sejumlah pohon keputusan. Setiap pohon dibentuk menggunakan sampel data dan sebagian fitur yang dipilih secara acak. Penggabungan tersebut mengurangi ketergantungan prediksi terhadap satu pohon dan memungkinkan model menangkap hubungan nonlinier serta interaksi antarfitur. Metode Random Forest mengacu pada pendekatan yang dikembangkan oleh (Breiman, 2001).

LightGBM dan XGBoost merupakan model *gradient boosting* yang membentuk pohon keputusan secara bertahap. Setiap pohon berikutnya diarahkan untuk memperbaiki kesalahan prediksi dari pohon sebelumnya. LightGBM dirancang untuk meningkatkan efisiensi pembentukan pohon pada data tabular, sedangkan XGBoost menggunakan fungsi objektif yang dilengkapi regularisasi untuk mengendalikan kompleksitas model. Penggunaan kedua model mengikuti pendekatan (Ke et al., 2017) untuk LightGBM serta (Chen & Guestrin, 2016)untuk XGBoost.

Seluruh model menjalani penalaan parameter menggunakan data pelatihan pada setiap *fold*. Ridge Regression menggunakan pencarian *grid* terhadap 25 nilai *alpha* yang tersebar secara logaritmik dari 10^{-4} sampai 10^{4}. Random Forest, LightGBM, dan XGBoost menggunakan Optuna dengan algoritma *Tree-structured Parzen Estimator*. Masing-masing model tersebut menjalani 50 *trial* pada setiap dataset dan *fold*. Konfigurasi yang menghasilkan RMSE validasi internal paling rendah dipilih sebagai konfigurasi terbaik. Proses tersebut menggunakan *random seed* 42 agar hasil pencarian parameter dapat direproduksi. Pembagian data untuk validasi internal dan penerapan *target-date purge* mengikuti prosedur yang telah dijelaskan pada Subbab 3.4. Konfigurasi dan ruang pencarian parameter setiap model ditunjukkan pada Tabel 3.6.

Tabel 3.6 Konfigurasi dan Penalaan Model

Model	Metode penalaan	Parameter dan ruang pencarian	Jumlah pencarian
Ridge Regression	Pencarian *grid*	*Alpha*: 10^{-4} sampai 10^{4} pada skala logaritmik	25 nilai
Random Forest	Optuna TPE	*n_estimators*: 150–700, *max_depth*: 3–20, *min_samples_split*: 2–40, *min_samples_leaf*: 1–30, *max_features*: 0,3–1,0, dan *max_samples*: 0,65–1,0	50 *trial*
LightGBM	Optuna TPE	*n_estimators*: 150–700, *learning_rate*: 0,01–0,08, *num_leaves*: 15–95, *max_depth*: 3–10, *min_child_samples*: 10–80, *subsample*: 0,65–1,0, *colsample_bytree*: 0,65–1,0, *reg_alpha*: 10^{-4}–5, dan *reg_lambda*: 10^{-4}–5	50 *trial*
XGBoost	Optuna TPE	*n_estimators*: 150–700, *learning_rate*: 0,01–0,08, *max_depth*: 2–8, *min_child_weight*: 1–12, *subsample*: 0,65–1,0, *colsample_bytree*: 0,65–1,0, *reg_alpha*: 10^{-4}–5, dan *reg_lambda*: 10^{-4}–8	50 *trial*

Penalaan parameter pada ketiga model berbasis pohon dilakukan menggunakan Optuna karena jumlah kombinasi parameter yang mungkin diuji relatif besar. Optuna memilih kombinasi parameter secara bertahap berdasarkan hasil *trial* sebelumnya sehingga pencarian dapat diarahkan pada bagian ruang parameter yang lebih menjanjikan. Setelah konfigurasi terbaik diperoleh, model dilatih kembali menggunakan seluruh data pelatihan pada *fold* yang bersangkutan sebelum menghasilkan prediksi pada data uji. Data uji tidak digunakan dalam pemilihan parameter.

Selain keempat model tersebut, penelitian ini menggunakan *no-change forecast* dan *persistence forecast* sebagai metode pembanding sederhana. Kedua metode ini tidak melakukan proses pembelajaran maupun penalaan parameter. *No-change forecast* menghasilkan prediksi *return* sebesar nol pada setiap periode pengujian. Prediksi tersebut dirumuskan sebagai berikut.

$$\hat{y}_{\text{no-change,t}} = 0 \tag{3.4}$$

Keterangan:

$\hat{y}_{\text{no-change,t}}$ = prediksi *return* dari *no-change forecast* pada waktu t.

Persistence forecast menggunakan *return* IHSG selama 20 sesi perdagangan sebelumnya sebagai prediksi untuk 20 sesi perdagangan berikutnya. Prediksi tersebut dihitung menggunakan persamaan berikut.

$$\hat{y}_{\text{persistence,t}} = \frac{\text{Close}_t}{\text{Close}_{t-20}} - 1 \quad (3.5)$$

Keterangan:
$\hat{y}_{\text{persistence,t}}$ = prediksi *return* dari *persistence forecast* pada waktu t.
Close_t = harga penutupan IHSG pada sesi ketika prediksi dibuat.
Close_{t-20} = harga penutupan IHSG 20 sesi perdagangan sebelumnya.

Kedua metode pembanding digunakan sebagai acuan dasar dalam mengevaluasi kemampuan prediksi keempat model. Perbandingan performa model dengan *no-change forecast* dan *persistence forecast* selanjutnya dilakukan menggunakan metrik evaluasi yang dijelaskan pada Subbab 3.6.

3.6 Metrik Evaluasi

Performa prediksi dalam penelitian ini diukur menggunakan *Root Mean Squared Error* (RMSE), *Mean Absolute Error* (MAE), dan *Directional Accuracy* (DA). RMSE digunakan sebagai metrik utama untuk menilai besar galat prediksi, sedangkan MAE dan DA digunakan sebagai metrik pendukung. RMSE dan MAE mengukur selisih antara nilai *return* aktual dan nilai prediksi, sedangkan DA mengukur kesesuaian arah pergerakan antara keduanya.

RMSE mengukur besar galat prediksi dengan memberikan pengaruh yang lebih besar terhadap galat bernilai besar. Hal tersebut terjadi karena selisih antara nilai aktual dan prediksi dikuadratkan sebelum dihitung rata-ratanya. Dengan demikian, beberapa prediksi yang meleset jauh akan meningkatkan nilai RMSE secara lebih besar. Nilai RMSE yang lebih kecil menunjukkan bahwa hasil prediksi model lebih dekat dengan nilai aktual. RMSE dihitung menggunakan persamaan berikut.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3.6)$$

Keterangan:
RMSE = *Root Mean Squared Error*.
n = jumlah observasi.
y_i = nilai *return* aktual pada observasi ke-i.
$\hat{y}_i$ = nilai *return* prediksi pada observasi ke-i.

MAE mengukur rata-rata nilai absolut dari selisih antara *return* aktual dan prediksi. Berbeda dengan RMSE, MAE tidak memperbesar pengaruh galat melalui proses pengkuadratan sehingga nilainya dapat dibaca sebagai rata-rata kesalahan absolut model.

Semakin rendah nilai MAE, semakin kecil rata-rata kesalahan absolut tersebut. MAE dihitung menggunakan persamaan berikut.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.7}$$

Keterangan:
MAE = *Mean Absolute Error*.
n = jumlah observasi.
y_i = nilai *return* aktual pada observasi ke-i.
$\hat{y}_i$ = nilai *return* prediksi pada observasi ke-i.

Selain mengukur besar galat, penelitian ini menggunakan *Directional Accuracy* untuk mengetahui kesesuaian arah antara *return* aktual dan hasil prediksi. Nilai *return* dikelompokkan menjadi arah positif dan nonpositif. *Return* dinyatakan positif apabila nilainya lebih besar dari nol, sedangkan nilai yang sama dengan atau lebih kecil dari nol dinyatakan nonpositif. Prediksi dianggap benar apabila *return* aktual dan prediksi berada pada kelompok arah yang sama.

Directional Accuracy dihitung sebagai persentase jumlah prediksi dengan arah yang benar terhadap seluruh observasi. Nilai DA yang lebih tinggi menunjukkan bahwa model lebih sering menghasilkan arah prediksi yang sesuai dengan arah *return* aktual. Perhitungan DA menggunakan persamaan berikut.

$$\mathrm{DA} = \frac{N_{benar}}{n} \times 100\% \tag{3.8}$$

Keterangan:
DA = *Directional Accuracy*.
N_{benar} = jumlah prediksi dengan arah yang benar.
n = jumlah observasi.

Nilai DA tidak digunakan untuk menggantikan RMSE dan MAE karena ketepatan arah tidak menunjukkan besar kesalahan prediksi. Model dapat menghasilkan arah prediksi yang benar, tetapi nilai prediksinya masih berbeda jauh dari *return* aktual. Sebaliknya, model dapat memiliki galat yang relatif kecil, tetapi menghasilkan arah yang salah ketika nilai aktual berada di sekitar nol. Oleh karena itu, ketiga metrik digunakan bersama agar evaluasi mempertimbangkan besar galat dan ketepatan arah prediksi.

Penelitian ini tidak menggunakan *Mean Absolute Percentage Error* (MAPE) sebagai metrik utama karena target *return* dapat bernilai nol, mendekati nol, atau negatif. Pembagian galat dengan nilai aktual yang mendekati nol dapat menghasilkan nilai MAPE yang sangat

besar dan sulit ditafsirkan. MAPE juga tidak dapat dihitung ketika nilai aktual sama dengan nol sehingga kurang sesuai untuk target dalam penelitian ini (Kolassa, 2026). *R-squared* juga tidak digunakan sebagai metrik utama karena pada prediksi *return* secara *out-of-sample*, nilainya dapat sangat kecil atau negatif ketika kemampuan prediksi model lemah. RMSE, MAE, dan DA dipilih karena memberikan ukuran yang lebih langsung terhadap besar galat dan kesesuaian arah prediksi (Cao, 2025).

3.7 Analisis Stabilitas Performa

Analisis stabilitas performa dilakukan untuk melihat kemampuan model mempertahankan performanya pada beberapa periode dan kondisi pasar yang berbeda. Model yang memiliki rata-rata galat rendah belum tentu menunjukkan performa yang stabil apabila galatnya berubah secara besar dari satu *fold* ke *fold* berikutnya. Oleh karena itu, stabilitas tidak hanya dinilai berdasarkan rata-rata RMSE dan MAE, tetapi juga berdasarkan variasi galat antar-*fold*, galat terbesar, performa pada setiap rezim volatilitas, dan perbandingan dengan metode pembanding sederhana. Seluruh ukuran tersebut dihitung pada tingkat kombinasi model dan dataset dengan menggunakan enam *fold* pengujian yang sama.

Variasi galat antar-*fold* diukur menggunakan standar deviasi RMSE dan MAE. Standar deviasi menunjukkan seberapa jauh nilai galat pada setiap *fold* menyimpang dari rata-ratanya. Nilai standar deviasi yang lebih kecil menunjukkan bahwa galat model relatif konsisten antar-*fold*, sedangkan nilai yang lebih besar menunjukkan adanya perubahan performa yang lebih tinggi. Standar deviasi galat dihitung menggunakan persamaan berikut.

$$\mathrm{SD_m} = \sqrt{\frac{1}{\mathrm{K}-1}\sum_{\mathrm{k}=1}^{\mathrm{K}}(\mathrm{E_{m,k}} - \overline{\mathrm{E}}_{\mathrm{m}})^2} \tag{3.9}$$

Keterangan:
$\mathrm{SD_m}$ = standar deviasi galat model ke-m.
K = jumlah *fold* pengujian.
$\mathrm{E_{m,k}}$ = nilai RMSE atau MAE model ke-m pada *fold* ke-k.
$\overline{\mathrm{E}}_{\mathrm{m}}$ = rata-rata RMSE atau MAE model ke-m dari seluruh *fold*.

Standar deviasi kemudian digunakan untuk menghitung *coefficient of variation* (CV). Ukuran ini membandingkan standar deviasi dengan rata-rata galat sehingga variasi performa dapat dibaca secara relatif. Penggunaan CV membantu membandingkan kestabilan model yang memiliki rata-rata RMSE atau MAE berbeda. Nilai CV yang lebih kecil menunjukkan bahwa

perubahan galat antar-*fold* lebih rendah dibandingkan rata-rata galat model tersebut. CV dihitung menggunakan persamaan berikut.

$$\mathrm{CV_m} = \frac{\mathrm{SD_m}}{\overline{\mathrm{E}}_\mathrm{m}} \tag{3.10}$$

Keterangan:
$\mathrm{CV_m}$ = *coefficient of variation* model ke-m.
$\mathrm{SD_m}$ = standar deviasi galat model ke-m.
$\overline{\mathrm{E}}_\mathrm{m}$ = rata-rata galat model ke-m.

Selain standar deviasi dan CV, penelitian ini menggunakan *worst-fold RMSE* dan *worst-fold MAE*. Kedua ukuran tersebut merupakan nilai RMSE dan MAE terbesar yang dihasilkan model selama enam *fold* pengujian. Ukuran ini digunakan untuk melihat kondisi ketika model menghasilkan galat paling besar. Untuk *Directional Accuracy*, stabilitas dibaca menggunakan nilai rata-rata, standar deviasi, dan nilai terendah antar-*fold* sebagai ukuran pendukung.

Rezim volatilitas dibentuk menggunakan *Markov Switching* dengan tiga *state*. Model tersebut diterapkan pada *return* harian IHSG dengan rata-rata konstan dan varians yang dapat berubah sesuai dengan *state* yang sedang berlangsung. Variabel $\mathrm{S_t}$ menunjukkan *state* volatilitas pada waktu t dan mengikuti proses Markov, yaitu probabilitas *state* pada waktu t ditentukan oleh *state* pada waktu sebelumnya (Hamilton, 1989). Bentuk model yang digunakan ditunjukkan pada persamaan berikut.

$$\mathrm{r_t} = \mu + \varepsilon_\mathrm{t}, \varepsilon_\mathrm{t} \sim \mathrm{N}(0, \sigma^2(\mathrm{S_t})), \mathrm{S_t} \in \{1,2,3\} \tag{3.11}$$

Probabilitas perpindahan dari *state* i menuju *state* j dirumuskan sebagai berikut.

$$\mathrm{p_{ij}} = \mathrm{P}(\mathrm{S_t} = \mathrm{j} \mid \mathrm{S_{t-1}} = \mathrm{i}) \tag{3.12}$$

Keterangan:
$\mathrm{r_t}$ = *return* harian IHSG pada waktu t.
μ = rata-rata *return* yang bersifat konstan.
$\sigma^2(\mathrm{S_t})$ = varians *return* pada *state* yang berlangsung.
$\mathrm{p_{ij}}$ = probabilitas perpindahan dari *state* i menuju *state* j.

Penentuan rezim melalui *Markov Switching* tidak menggunakan nilai *threshold* yang ditetapkan sebelum estimasi. Model terlebih dahulu mengestimasi varians dan probabilitas dari setiap *state*. Ketiga *state* kemudian diurutkan berdasarkan nilai variansnya. *State* dengan varians terkecil diberi label volatilitas rendah, *state* berikutnya diberi label volatilitas sedang,

dan *state* dengan varians terbesar diberi label volatilitas tinggi. Konfigurasi Markov Switching yang digunakan ditunjukkan pada Tabel 3.7.

Tabel 3.7 Konfigurasi Markov Switching

Komponen	Ketentuan
Variabel masukan	*Return* harian IHSG
Jumlah *state*	Tiga *state*
Parameter yang berubah	Varians
Parameter rata-rata	Konstan pada seluruh *state*
Data estimasi	Data harian sebelum awal periode pengujian pada setiap *fold*
Minimum data pelatihan	500 observasi harian
Maksimum iterasi	500 iterasi
Pengulangan pencarian	10 kali
Metode penetapan *state*	*Filtered probability argmax*
Penamaan rezim	Diurutkan berdasarkan varians setiap *state*

Estimasi *Markov Switching* dilakukan secara terpisah pada setiap *fold* menggunakan data sebelum awal periode pengujian. Setelah parameter diperoleh, *filtered probability* dihitung secara berurutan pada periode pengujian. Perhitungan tersebut hanya menggunakan informasi yang telah tersedia sampai tanggal observasi sehingga tidak menggunakan data dari tanggal berikutnya. Setiap tanggal pengujian ditempatkan pada *state* dengan *filtered probability* tertinggi. Label rezim pada tanggal yang sama digunakan untuk mengevaluasi hasil prediksi Dataset A, Dataset B, dan Dataset C agar perbandingan antardataset dilakukan pada kondisi pasar yang sama.

RMSE, MAE, dan DA selanjutnya dihitung untuk setiap rezim. Perbandingan formal hanya dilakukan apabila suatu rezim memiliki sedikitnya 10 observasi pengujian. Rezim dengan jumlah observasi di bawah batas tersebut tetap ditampilkan sebagai hasil deskriptif, tetapi tidak digunakan untuk menghitung *low-to-high degradation*, menentukan *worst regime* formal, atau mendukung kesimpulan stabilitas antarrezim. *Low-to-high degradation* menunjukkan persentase perubahan galat dari rezim volatilitas rendah menuju rezim volatilitas tinggi.

Stabilitas performa juga dinilai melalui *fold win rate* terhadap *no-change forecast* dan *persistence forecast*. Model dinyatakan mengungguli metode pembanding pada suatu *fold* apabila RMSE model lebih rendah daripada RMSE metode pembanding pada tahun pengujian yang sama. *Fold win rate* dihitung sebagai proporsi jumlah *fold* yang dimenangkan model terhadap enam *fold* pengujian. Ukuran ini digunakan agar pemilihan model tidak hanya didasarkan pada perbandingan antarmodel, tetapi juga mempertimbangkan kemampuannya mengungguli prediksi sederhana. Seluruh indikator stabilitas yang digunakan dirangkum pada Tabel 3.8.

Tabel 3.8 Ringkasan Indikator Stabilitas Performa Model

Indikator	**Fungsi**
Rata-rata RMSE dan MAE	Menilai besar galat model secara umum
Standar deviasi RMSE dan MAE	Menilai perubahan galat antar-*fold*
Coefficient of variation	Menilai variasi galat secara relatif terhadap rata-ratanya
Worst-fold RMSE dan *worst-fold MAE*	Menunjukkan galat terbesar selama pengujian
Rata-rata, standar deviasi, dan nilai terendah DA	Menilai ketepatan arah serta perubahannya antar-*fold*
Fold win rate terhadap metode pembanding	Menilai konsistensi model dalam mengungguli prediksi sederhana
RMSE, MAE, dan DA per rezim	Menilai performa pada kondisi volatilitas berbeda
Low-to-high degradation	Menilai perubahan galat dari volatilitas rendah menuju tinggi
Worst-regime RMSE dan *worst-regime MAE*	Menunjukkan rezim dengan galat terbesar

Penentuan model yang stabil dilakukan dengan membaca seluruh indikator secara bersama-sama. Model dengan rata-rata galat rendah belum langsung dinyatakan stabil apabila memiliki standar deviasi, CV, atau galat terburuk yang tinggi. Sebaliknya, model dengan variasi galat rendah tetap perlu dibandingkan berdasarkan tingkat galat dan kemampuannya

mengungguli metode pembanding. Dengan demikian, hasil analisis stabilitas tidak bergantung pada satu ukuran performa saja.

3.8 Uji Diebold-Mariano

Uji Diebold-Mariano digunakan untuk menilai apakah dua model memiliki perbedaan akurasi prediksi yang signifikan secara statistik (Diebold & Mariano, 2002). dilakukan terhadap pasangan prediksi pada dataset, tanggal, dan *fold* yang sama. Dengan pasangan tersebut, perbedaan galat tidak dipengaruhi oleh penggunaan observasi uji yang berbeda. Hipotesis nol menyatakan bahwa nilai harapan selisih *loss* kedua model sama dengan nol.

Misalkan $e_{a,t}$ dan $e_{b,t}$ merupakan galat prediksi model a dan model b pada waktu t. Selisih *loss* dihitung dengan mengurangkan *loss* model b dari *loss* model a sebagaimana ditunjukkan pada persamaan berikut.

$$d_t = L(e_{a,t}) - L(e_{b,t}) \tag{3.13}$$

Pengujian dilakukan menggunakan *squared error* dan *absolute error*. Kedua fungsi *loss* tersebut selaras dengan RMSE dan MAE yang digunakan dalam evaluasi performa. Bentuk fungsi *loss* dirumuskan sebagai berikut.

$$L_{SE}(e_t) = e_t^2, L_{AE}(e_t) = |e_t| \tag{3.14}$$

Statistik DM dihitung dari rata-rata selisih *loss* dan estimasi varians jangka panjangnya. Estimasi tersebut menggunakan pendekatan *heteroskedasticity and autocorrelation consistent* (HAC) dengan bobot Bartlett untuk mengakomodasi kemungkinan autokorelasi pada selisih *loss* (Newey & West, 1987). Perhitungan statistik DM dan varians jangka panjang ditunjukkan pada persamaan berikut.

$$DM = \frac{\bar{d}}{\sqrt{\frac{LRV}{n}}} \tag{3.15}$$

$$LRV = \gamma_0 + 2\sum_{l=1}^{q}\left[1-\frac{l}{q+1}\right]\gamma_l \tag{3.16}$$

Keterangan:

$\bar{d}$ = rata-rata selisih *loss*.
n = jumlah pasangan observasi.
LRV = estimasi varians jangka panjang selisih *loss*.
γ_l = autokovarians selisih *loss* pada *lag* ke-l.
q = jumlah *lag* HAC.

Nilai statistik DM positif menunjukkan bahwa model b memiliki *loss* lebih rendah, sedangkan nilai negatif menunjukkan bahwa model a memiliki *loss* lebih rendah. Pengujian menggunakan hipotesis dua arah dengan tingkat signifikansi 5%. Konfigurasi utama menggunakan HAC *lag* 1 karena prediksi dievaluasi dalam frekuensi bulanan. *Sensitivity test* dilakukan pada *lag* 0, 1, 2, dan 19. *Lag* 19 diperlakukan sebagai pemeriksaan konservatif dan bukan konfigurasi utama karena observasi galat tidak berfrekuensi harian.

Perbandingan berpasangan menghasilkan beberapa nilai p untuk setiap kelompok pengujian. Oleh karena itu, nilai p disesuaikan menggunakan koreksi Holm agar risiko kesimpulan signifikan akibat banyaknya perbandingan dapat dikendalikan (Holm, 1979). Kesimpulan utama didasarkan pada nilai p setelah koreksi Holm, sedangkan hasil pada setiap *lag* digunakan untuk melihat ketahanan kesimpulan terhadap pilihan panjang *lag*. Kelompok pengujian dibedakan menjadi perbandingan antarmodel pada dataset yang sama dan perbandingan antardataset pada setiap pasangan *ablation*, sehingga koreksi Holm diterapkan secara terpisah pada masing-masing kelompok dan setiap fungsi *loss*. Konfigurasi uji Diebold-Mariano ditunjukkan pada Tabel 3.9.

Tabel 3.9 Konfigurasi Uji Diebold-Mariano

Komponen	Ketentuan
Pasangan observasi	Dataset, tanggal, dan *fold* yang sama
Fungsi *loss*	*Squared error* dan *absolute error*
Estimasi varians	HAC dengan bobot Bartlett
Konfigurasi utama	HAC *lag* 1
Sensitivity test	HAC *lag* 0, 1, 2, dan 19
Hipotesis	Dua arah dengan tingkat signifikansi 5%
Koreksi perbandingan	Koreksi Holm pada setiap kelompok dan fungsi *loss*

3.9 *Ablation Study*

Ablation study digunakan untuk menilai perubahan kemampuan prediksi setelah kelompok fitur ditambahkan secara bertahap. Perbandingan dilakukan menggunakan model, tanggal evaluasi, *fold*, dan target yang sama agar perubahan hasil berkaitan dengan perbedaan susunan fitur. Analisis ini melengkapi perbandingan utama Dataset A, Dataset B, dan Dataset C dengan memisahkan kontribusi prediktif indikator domestik, indikator global, dan volatilitas GARCH. Susunan C tanpa GARCH memuat 65 fitur, yaitu seluruh fitur Dataset C tanpa fitur volatilitas GARCH. Rancangan *ablation study* ditunjukkan pada Tabel 3.10.

Tabel 3.10 Rancangan Ablation Study

Perbandingan	Kelompok fitur yang ditambahkan	Tujuan
Dataset A ke Dataset B	Indikator ekonomi domestik	Menilai perubahan setelah penambahan informasi domestik
Dataset B ke C tanpa GARCH	Indikator pasar global	Menilai perubahan setelah penambahan informasi global
C tanpa GARCH ke Dataset C	Volatilitas GARCH	Menilai perubahan setelah penambahan volatilitas GARCH

Susunan C tanpa GARCH hanya digunakan sebagai tampilan evaluasi sementara dalam *ablation study* dan tidak diperlakukan sebagai dataset utama keempat. Perubahan setiap metrik dihitung sebagai nilai setelah penambahan fitur dikurangi nilai sebelum penambahan fitur dengan rumus berikut.

$$\Delta \mathrm{M} = \mathrm{M}_{\mathrm{setelah}} - \mathrm{M}_{\mathrm{sebelum}} \tag{3.17}$$

Keterangan:
Δ_{M} = perubahan nilai metrik M setelah suatu kelompok fitur ditambahkan.
$\mathrm{M}_{\mathrm{setelah}}$ = nilai metrik sesudah kelompok fitur ditambahkan.
$\mathrm{M}_{\mathrm{sebelum}}$ = nilai metrik sebelum kelompok fitur ditambahkan.

Nilai delta negatif pada RMSE, MAE, standar deviasi galat, dan galat terburuk menunjukkan perbaikan setelah penambahan fitur. Pada DA, nilai delta positif menunjukkan peningkatan ketepatan arah. Perbedaan galat prediksi pada setiap pasangan *ablation* juga diuji menggunakan uji Diebold-Mariano dengan prosedur yang sama seperti pada Subbab 3.8.

Analisis berdasarkan rezim hanya digunakan untuk rezim yang memenuhi sedikitnya 10 observasi sebagaimana ditetapkan pada Subbab 3.7. Hasil *ablation study* ditafsirkan sebagai kontribusi prediktif tambahan pada sampel pengujian. Hasil tersebut tidak menunjukkan hubungan sebab-akibat antara kelompok fitur dan *return* IHSG.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Skenario Eksperimen

Bab ini menyajikan hasil evaluasi model dalam memprediksi *return* IHSG 20 sesi perdagangan ke depan. Hasil yang dibahas berasal dari eksperimen utama yang menggunakan tiga dataset, empat model prediksi, dua metode pembanding sederhana, dan evaluasi berbasis *expanding walk-forward validation*. Pembahasan difokuskan pada besar galat prediksi, stabilitas performa model pada beberapa kondisi pengujian, serta kekuatan bukti statistik yang mendukung perbedaan akurasi prediksi.

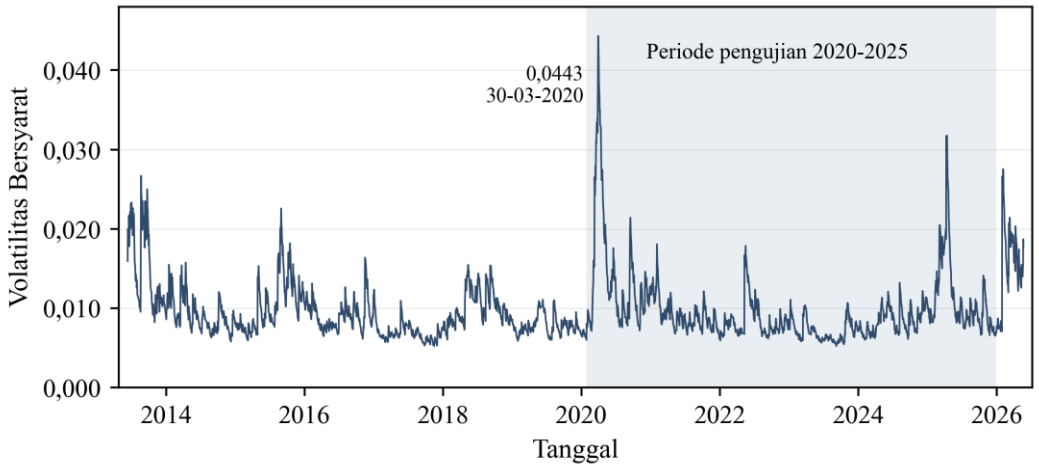
Dataset yang digunakan terdiri dari Dataset A, Dataset B, dan Dataset C sebagaimana dijelaskan pada Bab 3. Ketiga dataset dievaluasi pada periode uji yang sama, yaitu enam *fold* tahunan dari 2020 sampai dengan 2025 dengan 72 observasi bulanan untuk setiap kombinasi dataset dan model. Tanggal prediksi, target, dan pembagian *fold* identik pada ketiga dataset sehingga perbedaan hasil dapat dibaca sebagai perbedaan susunan fitur, bukan sebagai perbedaan cakupan periode pengujian.

Evaluasi model menggunakan RMSE sebagai metrik utama, sedangkan MAE dan *directional accuracy* digunakan sebagai metrik pendukung. Selain nilai rata-rata, hasil dibaca berdasarkan variasi galat antar-*fold*, performa tahunan, dan rezim volatilitas rendah, sedang, dan tinggi. Perbedaan akurasi antarmodel dan antardataset selanjutnya diuji menggunakan Diebold-Mariano dengan koreksi Holm, kemudian dilengkapi *ablation study*. Dengan demikian, pembahasan Bab 4 tidak hanya menampilkan model dengan galat rata-rata paling kecil, tetapi juga menunjukkan perubahan performa model pada periode dan kondisi pasar yang berbeda beserta kekuatan bukti yang mendukungnya.

Fitur volatilitas pada Dataset C berasal dari estimasi GARCH yang diperiksa pada setiap pembaruan parameter. Dari 149 estimasi sepanjang periode data, seluruhnya konvergen dan memenuhi syarat validitas parameter sehingga tidak ada estimasi yang perlu diulang menggunakan distribusi normal. Rata-rata nilai alpha sebesar 0,1145 dan beta sebesar 0,8624 dengan *persistence* rata-rata 0,9768 serta nilai tertinggi 0,9924, sehingga syarat jumlah alpha dan beta kurang dari satu terpenuhi pada seluruh estimasi. Diagnostik pada *lag* 10 tidak menemukan penolakan Ljung-Box pada kuadrat residual terstandarisasi maupun penolakan ARCH-LM, sehingga persamaan varians dinilai memadai. Uji Ljung-Box pada residual terstandarisasi menolak hipotesis nol pada 92 estimasi, yang menunjukkan bahwa autokorelasi

pada persamaan rata-rata belum sepenuhnya tertangkap oleh spesifikasi rata-rata konstan. Kondisi tersebut dicatat sebagai keterbatasan spesifikasi rata-rata dan dibaca bersama hasil diagnostik varians yang tidak menunjukkan sisa efek ARCH.

Sebaran volatilitas bersyarat yang dihasilkan estimasi tersebut ditunjukkan pada Gambar 4.1.



Gambar 4.1 Volatilitas Bersyarat GARCH(1,1) Periode 2013 sampai 2026

Berdasarkan Gambar 4.1, volatilitas bersyarat bergerak secara berkelompok, yaitu periode bergejolak cenderung diikuti periode bergejolak dan periode tenang diikuti periode tenang. Nilai tertinggi sebesar 0,044336 terjadi pada 30 Maret 2020 atau sekitar lima kali nilai tengah seluruh periode yang sebesar 0,008547. Rata-rata tahunan tertinggi juga terjadi pada 2020 sebesar 0,013878, sedangkan yang terendah pada 2023 sebesar 0,007081. Pola pengelompokan tersebut mendukung penggunaan GARCH sebagai sumber fitur volatilitas dan Markov Switching sebagai dasar pembentukan rezim pada Subbab 3.7. Rentang gambar mencakup seluruh periode data yang tersedia, sedangkan evaluasi model pada bab ini terbatas pada enam *fold* pengujian 2020 sampai dengan 2025.

Seluruh hasil pada bab ini diperoleh dari eksperimen yang dijalankan menggunakan rancangan metodologi pada Bab 3. Pembahasan dibatasi pada evaluasi prediksi *return*, bukan pada pengujian profitabilitas atau keputusan investasi.

4.2 Ringkasan Performa Model pada Setiap Dataset

Evaluasi umum dilakukan untuk membandingkan performa empat model *machine learning* pada Dataset A, Dataset B, dan Dataset C. Perbandingan menggunakan *Mean* RMSE,

Mean MAE, dan *Worst-fold* RMSE, serta dilengkapi standar deviasi RMSE antar-*fold* dan *directional accuracy* sebagaimana ditetapkan pada Subbab 3.6 dan Subbab 3.7. Ringkasan performa seluruh model *machine learning* ditunjukkan pada Tabel 4.1.

Tabel 4.1 Performa Model

Dataset	**Model**	***Mean* RMSE**	***Mean* MAE**	***Worst-fold* RMSE**
A	Random Forest	0,044809	0,033772	0,106429
A	XGBoost	0,045360	0,034350	0,105896
A	Ridge Regression	0,045617	0,035363	0,107375
A	LightGBM	0,046240	0,035188	0,108595
B	Random Forest	0,044536	0,033355	0,106435
B	LightGBM	0,045749	0,034543	0,108595
B	XGBoost	0,048294	0,037560	0,109326
B	Ridge Regression	0,050440	0,038843	0,134241
C	Random Forest	0,044725	0,033693	0,106525
C	XGBoost	0,045295	0,034858	0,110052
C	Ridge Regression	0,045430	0,034887	0,107370
C	LightGBM	0,046476	0,035023	0,108595

Berdasarkan Tabel 4.1, Random Forest menghasilkan *Mean* RMSE paling rendah pada ketiga dataset. Pada Dataset A nilainya 0,044809, pada Dataset B turun menjadi 0,044536 yang merupakan nilai terendah di antara seluruh kombinasi dataset dan model, sedangkan pada Dataset C sebesar 0,044725. Peringkat tersebut bersifat deskriptif pada sampel evaluasi ini dan belum menunjukkan keunggulan yang didukung pengujian statistik. Pengujian tersebut dibahas pada Subbab 4.6.

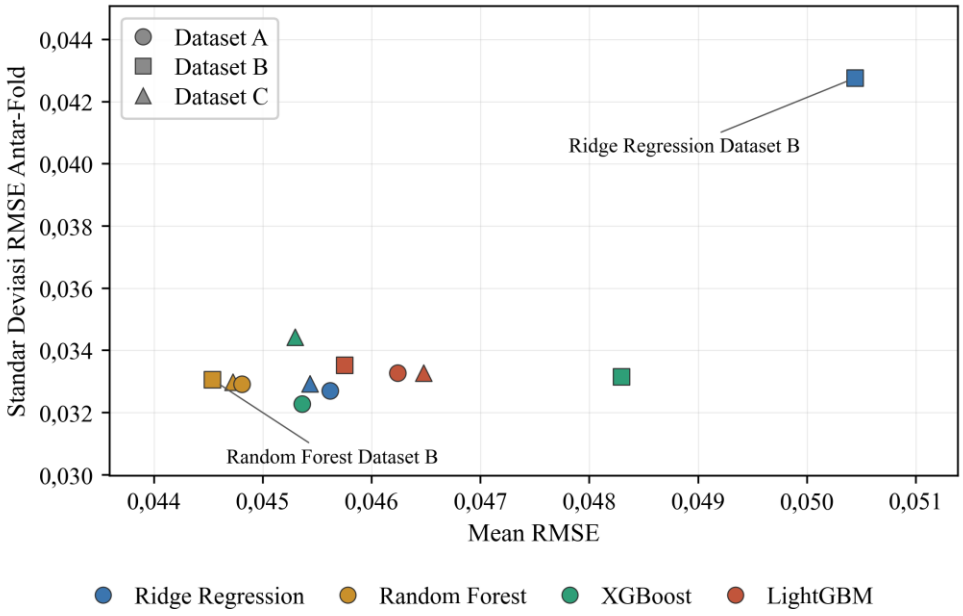
Perbandingan antardataset menunjukkan selisih yang kecil. Selisih *Mean* RMSE Random Forest antara Dataset B dan Dataset A hanya 0,000273, sedangkan antara Dataset B dan Dataset C hanya 0,000189. Karena ketiga dataset dievaluasi pada tanggal dan *fold* yang sama, selisih tersebut dapat dibaca sebagai perbedaan susunan fitur. Namun selisih itu jauh lebih kecil daripada standar deviasi galat antar-*fold*, sehingga urutan dataset belum dapat dianggap mantap hanya berdasarkan nilai rata-rata.

Model lain menghasilkan *Mean* RMSE yang berdekatan. XGBoost menempati urutan kedua pada Dataset A dan Dataset C, sedangkan LightGBM menempati urutan kedua pada Dataset B. Ridge Regression menghasilkan *Mean* RMSE tertinggi pada Dataset B, yaitu 0,050440, dengan standar deviasi 0,042761 yang juga paling besar di antara seluruh kombinasi.

Hasil tersebut menunjukkan bahwa model linear tidak selalu menghasilkan galat yang lebih stabil daripada model berbasis pohon pada rancangan evaluasi ini.

Nilai *Worst-fold* RMSE menunjukkan bahwa seluruh model masih memiliki periode pengujian dengan galat yang cukup besar. Pada Random Forest, *Worst-fold* RMSE berada pada 0,106429 untuk Dataset A, 0,106435 untuk Dataset B, dan 0,106525 untuk Dataset C, atau sekitar 2,4 kali *Mean* RMSE-nya. Standar deviasi RMSE antar-*fold* berkisar antara 0,032274 dan 0,042761, sehingga *coefficient of variation* seluruh kombinasi berada pada rentang 0,69 sampai 0,85. Artinya, variasi galat antar-*fold* berukuran hampir sebesar rata-ratanya sendiri pada seluruh model.

Hubungan antara galat rata-rata dan variasi galat antar-*fold* untuk kedua belas kombinasi ditunjukkan pada Gambar 4.2. Posisi yang semakin ke kiri berarti galat rata-rata lebih rendah dan posisi yang semakin ke bawah berarti variasi antar-*fold* lebih kecil.



Gambar 4.2 Mean RMSE terhadap Standar Deviasi RMSE Antar-Fold

Berdasarkan Gambar 4.2, sebelas dari dua belas kombinasi berkumpul pada wilayah yang sempit, yaitu *Mean* RMSE antara 0,044536 dan 0,048294 dengan standar deviasi antara 0,032274 dan 0,034428. Ridge Regression Dataset B terpisah jauh dari kelompok tersebut pada kedua sumbu sekaligus. Kombinasi selebihnya tidak membentuk pertukaran yang jelas antara galat rata-rata yang rendah dan variasi antar-*fold* yang kecil, sehingga tidak ada kombinasi yang

dapat disebut lebih akurat sekaligus lebih stabil. Pembacaan ini bersifat deskriptif dan kekuatan buktinya diuji pada Subbab 4.6.

Dua catatan perlu disampaikan sebelum angka kestabilan tersebut dibaca lebih jauh. LightGBM pada *fold* 2020 menghasilkan satu nilai prediksi yang sama untuk seluruh dua belas observasi, dan nilai tersebut identik pada ketiga dataset, sehingga *Worst-fold* RMSE LightGBM tercatat sama persis sebesar 0,108595 pada Dataset A, Dataset B, dan Dataset C. XGBoost pada *fold* 2022 menghasilkan pola serupa untuk Dataset A. Parameter yang terpilih pada *fold* tersebut juga identik untuk ketiga dataset, termasuk nilai regularisasi *reg_alpha* sebesar 1,175, yang menunjukkan bahwa model tidak menggunakan fitur pada *fold* itu. Standar deviasi dan *Worst-fold* RMSE pada kombinasi yang terdampak karena itu belum sepenuhnya mencerminkan perilaku model terhadap fitur. Angka tersebut tetap dilaporkan apa adanya dan dicatat sebagai keterbatasan penelitian.

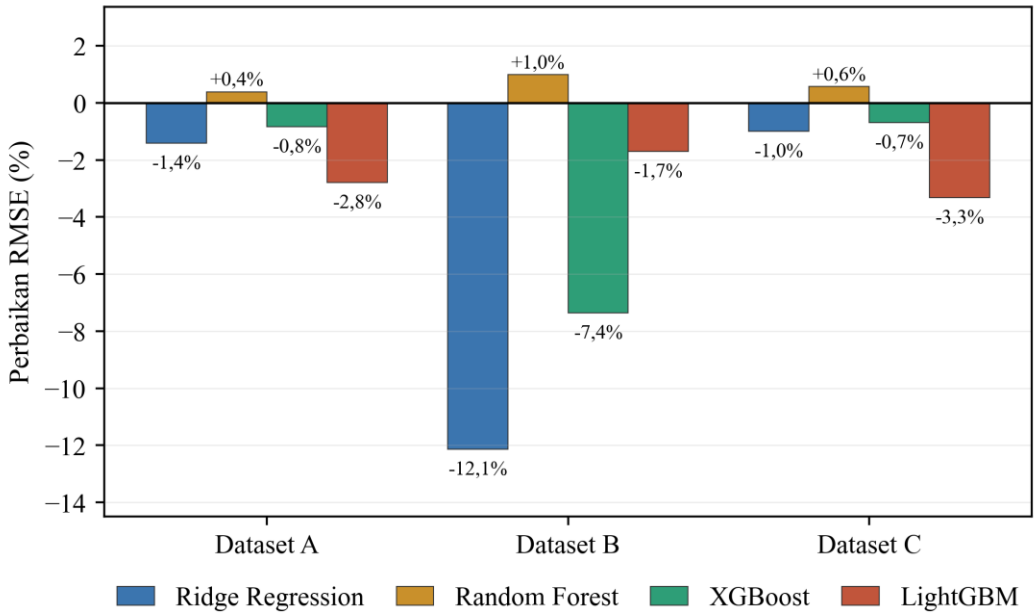
Directional accuracy berada pada rentang 50,00% sampai 61,11%. Nilai tertinggi diperoleh Random Forest pada Dataset B, sedangkan Ridge Regression pada Dataset A dan Dataset B berada tepat pada 50,00%. Nilai tersebut dibaca sebagai metrik pendukung dan tidak menggantikan RMSE, karena ketepatan arah tidak menunjukkan besar galat prediksi.

Oleh karena itu, hasil evaluasi ini perlu dilanjutkan dengan pembacaan terhadap metode pembanding sederhana, performa tahunan, dan rezim volatilitas, kemudian diuji kekuatan buktinya melalui Diebold-Mariano dan *ablation study* pada subbab berikutnya.

4.3 Perbandingan dengan Metode Pembanding Sederhana

Perbandingan dengan metode pembanding sederhana digunakan untuk menilai apakah model *machine learning* memberikan nilai tambah dibandingkan aturan prediksi dasar. Pada evaluasi ini, *no-change forecast* menghasilkan *Mean* RMSE sebesar 0,044984 dan *persistence forecast* sebesar 0,055250. Kedua nilai tersebut berlaku sama untuk ketiga dataset karena keduanya hanya bergantung pada tanggal pengujian.

Perbaikan RMSE setiap model terhadap *no-change forecast* ditunjukkan pada Gambar 4.3. Nilai positif berarti model menghasilkan RMSE yang lebih rendah daripada metode pembanding.



Gambar 4.3 Perbaikan RMSE Setiap Model terhadap No-change Forecast

Berdasarkan Gambar 4.3, hanya tiga dari dua belas kombinasi dataset dan model yang menghasilkan RMSE lebih rendah daripada *no-change forecast*. Ketiganya adalah Random Forest, yaitu 0,996% pada Dataset B, 0,576% pada Dataset C, dan 0,389% pada Dataset A. Sembilan kombinasi lainnya menghasilkan RMSE yang lebih tinggi daripada metode pembanding tersebut, dengan selisih terbesar pada Ridge Regression Dataset B yang 12,129% lebih tinggi.

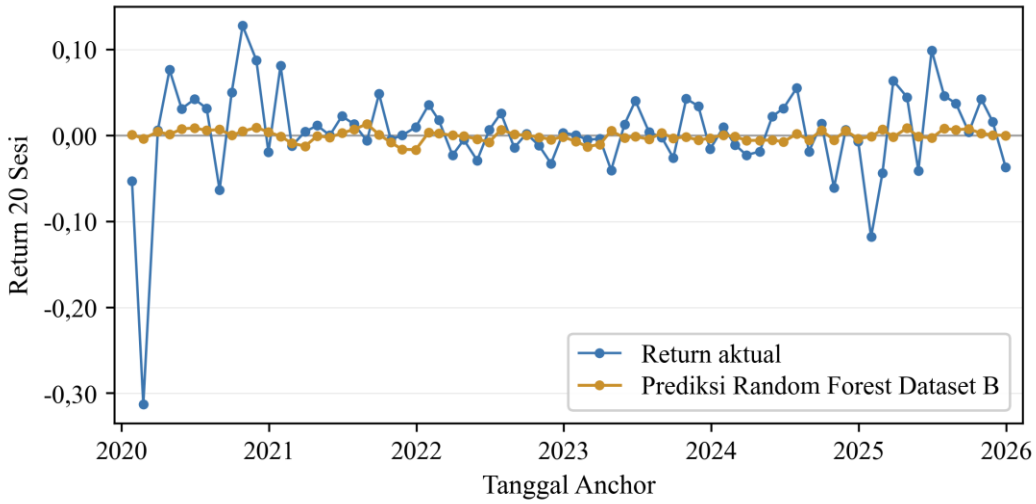
Terhadap *persistence forecast*, seluruh dua belas kombinasi menghasilkan RMSE yang lebih rendah. Perbaikannya berkisar antara 8,706% pada Ridge Regression Dataset B dan 19,392% pada Random Forest Dataset B. Perbedaan pola tersebut menunjukkan bahwa kedua metode pembanding memiliki tingkat kesulitan yang tidak sama, dan *no-change forecast* merupakan acuan yang jauh lebih ketat.

Selain rata-rata RMSE, perbandingan juga dibaca melalui *fold win rate*, yaitu proporsi *fold* ketika model menghasilkan RMSE lebih rendah daripada metode pembanding. Terhadap *no-change forecast*, Random Forest unggul pada empat dari enam *fold* di ketiga dataset, sedangkan model lain hanya unggul pada dua sampai tiga *fold*. Terhadap *persistence forecast*, hampir seluruh kombinasi unggul pada lima dari enam *fold*, kecuali XGBoost pada Dataset B yang unggul pada empat *fold*.

Pola tersebut menunjukkan bahwa keunggulan model terhadap *no-change forecast* tidak merata pada seluruh periode pengujian, bahkan pada kombinasi yang unggul berdasarkan rata-

rata. Random Forest Dataset B memperoleh perbaikan rata-rata sebesar 0,996%, tetapi tetap tertinggal pada dua *fold*.

Penyebab kecilnya selisih terhadap *no-change forecast* dapat dilihat dari sebaran prediksi model itu sendiri. Perbandingan antara *return* aktual dan prediksi Random Forest Dataset B pada seluruh 72 observasi pengujian ditunjukkan pada Gambar 4.4.



Gambar 4.4 Return Aktual dan Prediksi Random Forest Dataset B

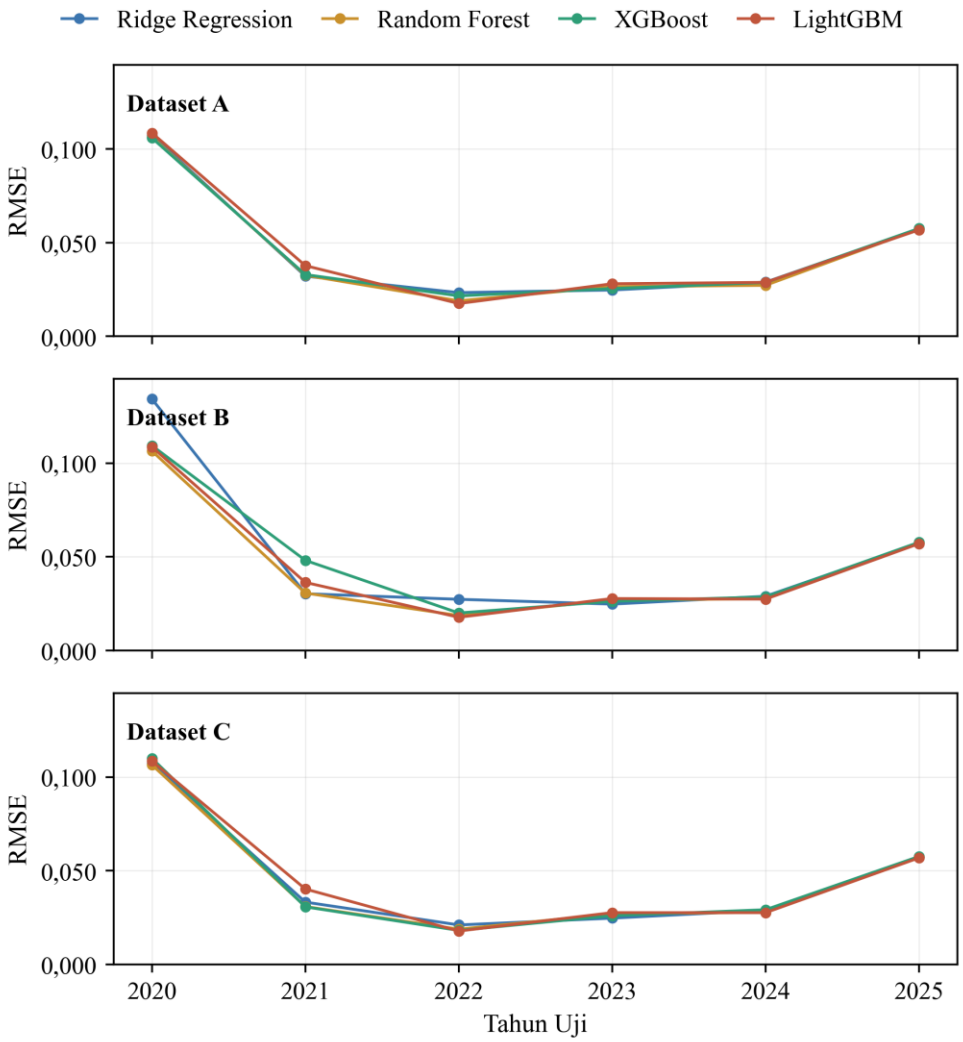
Berdasarkan Gambar 4.4, prediksi model bergerak pada rentang yang jauh lebih sempit daripada *return* aktual. Nilai prediksi berkisar antara −0,016509 dan 0,013494, sedangkan *return* aktual berkisar antara −0,312987 dan 0,127746. Standar deviasi prediksi sebesar 0,006118 atau sekitar sepersembilan standar deviasi *return* aktual yang sebesar 0,054778. Model karena itu hampir selalu memprediksi nilai yang mendekati nol dan jarang menyimpang jauh dari sana. Perilaku tersebut menjelaskan mengapa perbaikan terhadap *no-change forecast* yang selalu memprediksi nol hanya sebesar 0,996%, sedangkan perbaikan terhadap *persistence forecast* mencapai 19,392% karena metode itu mengulang perubahan sebelumnya yang berayun jauh lebih lebar.

Temuan tersebut menunjukkan bahwa prediksi *return* IHSG 20 sesi perdagangan ke depan memiliki sinyal prediktif yang terbatas. Model *machine learning* dapat menurunkan galat dibandingkan *persistence forecast*, tetapi perbaikannya terhadap *no-change forecast* sangat kecil dan hanya dicapai oleh satu jenis model. Dengan demikian, hasil penelitian ini tidak dibaca sebagai bukti bahwa model dapat memprediksi *return* secara kuat pada semua kondisi, melainkan sebagai evaluasi bahwa nilai tambah model terhadap metode pembanding

sederhana masih perlu dibaca secara hati-hati. Kekuatan bukti statistik atas perbedaan tersebut diuji pada Subbab 4.6.

4.4 Evaluasi Tahunan *Walk-Forward Validation*

Evaluasi tahunan dilakukan untuk melihat perubahan galat model dari satu periode uji ke periode uji berikutnya. Enam *fold* tahunan dari 2020 sampai dengan 2025 dievaluasi pada tanggal yang sama untuk ketiga dataset, sehingga perubahan galat dapat dibandingkan langsung antarmodel dan antardataset. RMSE setiap model pada setiap *fold* ditunjukkan pada Gambar 4.5.



Gambar 4.5 RMSE Tahunan Setiap Model pada Enam Fold Pengujian

Berdasarkan Gambar 4.5, galat prediksi berubah tajam antartahun. Seluruh model mencatat galat tertinggi pada 2020. Pada Random Forest Dataset B, RMSE 2020 sebesar

0,106435, sedangkan RMSE terendah terjadi pada 2022 sebesar 0,018589 atau kurang dari seperlima nilai tersebut. Sebelas dari dua belas kombinasi mencatat galat terendah pada 2022, sedangkan Ridge Regression Dataset B mencatatnya pada 2023.

Perbandingan per *fold* terhadap *no-change forecast* memperlihatkan pola yang tidak konsisten. Random Forest Dataset B menghasilkan RMSE lebih rendah pada 2020, 2022, 2024, dan 2025, tetapi lebih tinggi pada 2021 dan 2023. Keunggulan rata-rata sebesar 0,996% pada subbab sebelumnya karena itu terbentuk dari empat *fold* yang unggul dan dua *fold* yang tertinggal, bukan dari keunggulan yang merata.

Perubahan tersebut sejalan dengan standar deviasi RMSE antar-*fold* yang dilaporkan pada Subbab 4.2. Nilai standar deviasi yang hampir sebesar rata-ratanya sendiri berasal terutama dari selisih besar antara 2020 dan tahun-tahun berikutnya. Dengan demikian, nilai rata-rata RMSE tidak cukup untuk menjelaskan perilaku model karena rata-rata dapat menyembunyikan periode ketika galat meningkat tajam.

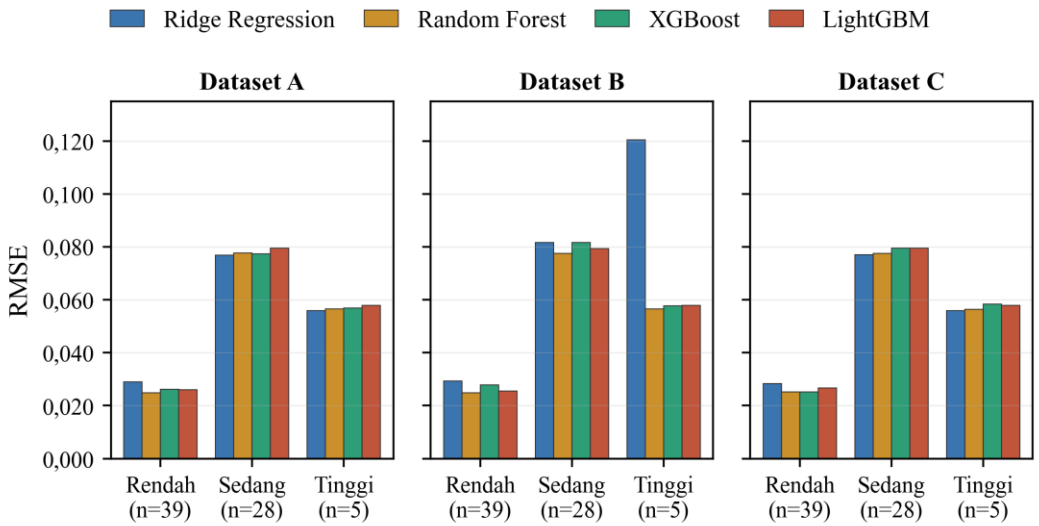
Dengan demikian, *expanding walk-forward validation* memberikan gambaran yang lebih lengkap dibandingkan evaluasi pada satu periode uji saja. Melalui evaluasi tahunan, penelitian ini dapat menunjukkan bahwa model dengan performa rata-rata terbaik tetap mengalami perubahan galat yang besar pada periode tertentu. Hasil ini mendukung fokus penelitian pada stabilitas performa model, yaitu melihat apakah galat model tetap rendah atau berubah ketika diuji pada beberapa periode waktu.

4.5 Evaluasi Berdasarkan Rezim Volatilitas

Evaluasi berdasarkan rezim volatilitas dilakukan untuk melihat perubahan galat model pada kondisi pasar rendah, sedang, dan tinggi. Rezim dibentuk menggunakan Markov Switching sebagaimana dijelaskan pada Subbab 3.7, dan label rezim yang sama diterapkan pada ketiga dataset karena tanggal pengujiannya identik. Pengelompokan ini digunakan setelah prediksi dibuat, sehingga hasilnya dibaca sebagai evaluasi performa, bukan sebagai tugas prediksi volatilitas.

Dari 72 observasi pengujian, 39 observasi dikelompokkan sebagai volatilitas rendah, 28 observasi sebagai volatilitas sedang, dan hanya 5 observasi sebagai volatilitas tinggi. Sebaran tersebut tidak merata antartahun. Seluruh observasi rezim tinggi berasal dari dua tahun saja, yaitu empat observasi pada 2020 dan satu observasi pada 2025, sedangkan seluruh observasi pada 2023 termasuk rezim rendah. Sesuai ketentuan pada Subbab 3.7, perbandingan formal hanya dilakukan pada rezim dengan sedikitnya 10 observasi, sehingga rezim tinggi ditampilkan

sebagai hasil deskriptif dan tidak digunakan untuk menghitung *low-to-high degradation* maupun menentukan rezim terburuk secara formal. RMSE setiap model pada setiap rezim ditunjukkan pada Gambar 4.6.



Gambar 4.6 RMSE Setiap Model pada Tiga Rezim Volatilitas

Berdasarkan Gambar 4.6, galat prediksi tidak meningkat secara berurutan dari rezim rendah ke rezim tinggi. Galat terbesar justru terjadi pada rezim sedang. Pada Random Forest Dataset B, RMSE rezim rendah sebesar 0,024937, rezim sedang 0,077499, dan rezim tinggi 0,056513, atau berturut-turut sekitar 3,1 kali dan 2,3 kali nilai rezim rendah. Pola yang sama terjadi pada sebelas dari dua belas kombinasi dataset dan model.

Peringkat model juga berbeda antarrezim. Pada rezim rendah, Random Forest Dataset A menghasilkan RMSE terendah sebesar 0,024915. Pada rezim sedang, Ridge Regression Dataset A menghasilkan RMSE terendah sebesar 0,076941. Perbedaan tersebut menunjukkan bahwa model yang unggul berdasarkan rata-rata keseluruhan belum tentu unggul pada setiap kondisi pasar.

Pada rezim tinggi, RMSE terendah dicapai Ridge Regression Dataset C sebesar 0,055938, sedangkan RMSE tertinggi dicapai Ridge Regression Dataset B sebesar 0,120544 atau lebih dari dua kali lipat kombinasi lainnya. Kombinasi tersebut juga menjadi satu-satunya pengecualian dari pola pada paragraf sebelumnya. Rentang yang sangat lebar itu terjadi pada kelompok yang hanya berisi lima observasi, sehingga satu prediksi yang meleset jauh dapat mengubah nilai RMSE secara besar. Karena itu, seluruh angka rezim tinggi dibaca sebagai hasil eksploratif dan tidak digunakan sebagai dasar kesimpulan stabilitas antarrezim.

Perbandingan terhadap metode pembanding sederhana pada setiap rezim juga tidak seragam. Pada rezim rendah, *no-change forecast* menghasilkan RMSE sebesar 0,025627 dan hanya lima dari dua belas kombinasi yang berada di bawahnya, tiga di antaranya Random Forest. Pada rezim sedang, *persistence forecast* justru menghasilkan RMSE terendah sebesar 0,072997, lebih rendah daripada seluruh model maupun *no-change forecast*. Temuan tersebut memperkuat pembacaan pada Subbab 4.3 bahwa nilai tambah model terhadap aturan prediksi sederhana masih terbatas.

Interpretasi pada bagian ini bersifat deskriptif. Penelitian ini tidak menyatakan bahwa volatilitas secara langsung menyebabkan galat model meningkat. Hasil yang dapat disimpulkan adalah bahwa besar galat prediksi berbeda antarkelompok volatilitas, dengan galat terbesar pada rezim sedang dan sebaran hasil yang sangat lebar pada rezim tinggi yang jumlah observasinya terbatas.

4.6 Uji Diebold-Mariano

Uji Diebold-Mariano digunakan untuk menilai apakah selisih akurasi prediksi yang terlihat pada subbab sebelumnya cukup kuat untuk disebut signifikan secara statistik. Pengujian dilakukan pada pasangan prediksi dengan dataset, tanggal, dan *fold* yang sama menggunakan prosedur pada Subbab 3.8. Pada *lag* utama HAC 1 terdapat 144 perbandingan, yang terdiri dari 120 perbandingan antarmodel dalam dataset yang sama dan 24 perbandingan antardataset pada pasangan *ablation*. Statistik DM yang bernilai positif menunjukkan bahwa anggota kedua pada pasangan memiliki *loss* lebih rendah.

Pada kelompok perbandingan antarmodel, 28 dari 120 perbandingan signifikan sebelum koreksi, tetapi tidak satu pun bertahan setelah koreksi Holm. Nilai p terkoreksi terkecil pada kelompok ini adalah 0,0767. Hasil untuk Random Forest Dataset B sebagai kombinasi dengan galat rata-rata terendah ditunjukkan pada Tabel 4.2.

Tabel 4.2 Uji Diebold-Mariano Random Forest Dataset B pada Lag HAC 1

Perbandingan	Fungsi *loss*	Statistik DM	Nilai p	Nilai p Holm
Random Forest – *No-change forecast*	*Squared error*	−1,509	0,1313	1,0000
Random Forest – *No-change forecast*	*Absolute error*	−0,431	0,6666	1,0000
Persistence forecast – Random Forest	*Squared error*	+1,215	0,2245	1,0000
Persistence forecast – Random Forest	*Absolute error*	+3,221	0,0013	0,0767
Random Forest – Ridge Regression	*Squared error*	−1,716	0,0862	1,0000
Random Forest – Ridge Regression	*Absolute error*	−2,031	0,0423	1,0000
Random Forest – XGBoost	*Squared error*	−2,050	0,0404	1,0000
Random Forest – XGBoost	*Absolute error*	−2,620	0,0088	0,4041
LightGBM – Random Forest	*Squared error*	+1,580	0,1142	1,0000
LightGBM – Random Forest	*Absolute error*	+1,769	0,0769	1,0000

Berdasarkan Tabel 4.2, selisih Random Forest Dataset B terhadap *no-change forecast* tidak signifikan pada kedua fungsi *loss*, bahkan sebelum koreksi. Nilai p sebesar 0,1313 pada *squared error* dan 0,6666 pada *absolute error* menunjukkan bahwa perbaikan rata-rata sebesar 0,996% pada Subbab 4.3 tidak dapat dibedakan dari variasi acak pada 72 observasi yang tersedia.

Perbandingan yang paling mendekati signifikan adalah selisih terhadap *persistence forecast* pada *absolute error*, dengan nilai p mentah 0,0013 yang berubah menjadi 0,0767 setelah koreksi Holm. Perbandingan terhadap XGBoost pada *absolute error* juga signifikan sebelum koreksi dengan nilai p 0,0088, tetapi menjadi 0,4041 setelah koreksi. Perubahan tersebut terjadi karena kelompok perbandingan antarmodel memuat 60 pasangan untuk setiap fungsi *loss*, sehingga koreksi Holm menuntut bukti yang jauh lebih kuat.

Jumlah hasil signifikan juga bergantung pada panjang *lag* HAC yang digunakan. Perbandingan pada keempat *lag* ditunjukkan pada Tabel 4.3.

Tabel 4.3 Sensitivitas Hasil Signifikan terhadap Panjang Lag HAC

Lag HAC	Signifikan mentah	Signifikan setelah Holm	Kelompok antarmodel	Kelompok *ablation*
0	35	2	0	2
1	33	1	0	1
2	29	10	9	1
19	23	14	14	0

Berdasarkan Tabel 4.3, jumlah hasil signifikan setelah koreksi justru bertambah pada *lag* yang lebih panjang, dari satu hasil pada *lag* 1 menjadi 14 hasil pada *lag* 19, dan seluruh tambahan tersebut berasal dari kelompok perbandingan antarmodel. Pola itu menunjukkan bahwa estimasi varians jangka panjang pada data ini sensitif terhadap panjang *lag* yang dipilih. Sesuai ketentuan pada Subbab 3.8, *lag* 1 tetap digunakan sebagai konfigurasi utama karena observasi dievaluasi dalam frekuensi bulanan. Hasil pada *lag* yang lebih panjang dilaporkan sebagai *sensitivity test* dan tidak dipilih sebagai dasar kesimpulan meskipun menghasilkan lebih banyak temuan signifikan.

Dengan demikian, peringkat model pada Subbab 4.2 tetap merupakan peringkat deskriptif. Pada *lag* utama, tidak ada perbandingan antarmodel maupun perbandingan terhadap metode pembanding sederhana yang bertahan setelah koreksi Holm. Random Forest Dataset B dapat disebut memiliki galat rata-rata terendah pada sampel evaluasi ini, tetapi tidak dapat disebut lebih akurat secara statistik daripada model lain.

4.7 Ablation Study

Ablation study digunakan untuk menilai perubahan kemampuan prediksi setelah kelompok fitur ditambahkan secara bertahap, sebagaimana dirancang pada Subbab 3.9. Perbandingan dilakukan pada model, tanggal evaluasi, *fold*, dan target yang sama sehingga perubahan hasil berkaitan dengan perbedaan susunan fitur. Perubahan *Mean* RMSE beserta hasil uji Diebold-Mariano pada setiap tahap ditunjukkan pada Tabel 4.4.

Tabel 4.4 Perubahan Mean RMSE dan Nilai p Holm Setiap Tahap Ablation

Tahap	**Model**	**Δ *Mean* RMSE**	**p Holm *squared***	**p Holm *absolute***
A ke B	Ridge Regression	+0,004824	0,3052	0,3331
A ke B	Random Forest	−0,000273	0,3052	0,3331
A ke B	XGBoost	+0,002934	0,3052	0,1200
A ke B	LightGBM	−0,000491	0,4243	0,3331
B ke C tanpa GARCH	Ridge Regression	−0,004998	0,2764	0,3857
B ke C tanpa GARCH	Random Forest	+0,000452	0,1175	0,0213
B ke C tanpa GARCH	XGBoost	+0,002776	0,2764	0,3857
B ke C tanpa GARCH	LightGBM	+0,000121	0,7513	0,5220
C tanpa GARCH ke C	Ridge Regression	−0,000012	0,7320	0,8878
C tanpa GARCH ke C	Random Forest	−0,000264	0,5597	0,1345
C tanpa GARCH ke C	XGBoost	−0,005775	0,3009	0,1345
C tanpa GARCH ke C	LightGBM	+0,000605	0,7320	0,8878

Berdasarkan Tabel 4.4, arah perubahan berbeda antarmodel pada setiap tahap. Penambahan indikator ekonomi domestik dari Dataset A ke Dataset B menurunkan *Mean* RMSE pada Random Forest sebesar 0,000273 dan pada LightGBM sebesar 0,000491, tetapi menaikkannya pada Ridge Regression sebesar 0,004824 dan pada XGBoost sebesar 0,002934. Tidak ada perbedaan pada tahap ini yang signifikan setelah koreksi Holm.

Penambahan indikator pasar global dari Dataset B ke C tanpa GARCH menghasilkan satu-satunya perbedaan yang bertahan setelah koreksi. Pada Random Forest, *absolute error* meningkat dan perbedaannya signifikan dengan nilai p terkoreksi 0,0213. Statistik DM bernilai negatif, yang berarti Dataset B memiliki *loss* lebih rendah. Dengan kata lain, penambahan indikator pasar global memperburuk galat absolut Random Forest, bukan memperbaikinya. Pada Ridge Regression tahap yang sama justru menurunkan *Mean* RMSE sebesar 0,004998, tetapi perbedaan tersebut tidak signifikan setelah koreksi.

Penambahan volatilitas GARCH dari C tanpa GARCH ke Dataset C menurunkan *Mean* RMSE pada XGBoost sebesar 0,005775, pada Random Forest sebesar 0,000264, dan pada Ridge Regression sebesar 0,000012, sedangkan pada LightGBM menaikkannya sebesar 0,000605. Perbedaan pada tahap ini signifikan sebelum koreksi untuk Random Forest dan XGBoost pada *absolute error*, dengan nilai p 0,0336 dan 0,0439, tetapi tidak ada yang bertahan setelah koreksi Holm.

Perubahan *directional accuracy* juga tidak seragam. Penambahan indikator domestik menaikkan *directional accuracy* Random Forest sebesar 4,17 poin persentase, sedangkan penambahan indikator global menurunkannya sebesar 2,78 poin persentase.

Hasil *ablation study* ditafsirkan sebagai kontribusi prediktif tambahan pada sampel pengujian ini. Hasil tersebut tidak menunjukkan hubungan sebab-akibat antara kelompok fitur dan *return* IHSG. Dari dua belas perbandingan pada tiga tahap, hanya satu yang bertahan setelah koreksi, dan arahnya menunjukkan penurunan kualitas prediksi. Karena itu, penambahan informasi di luar fitur teknikal belum dapat dinyatakan memberikan kontribusi prediktif yang konsisten.

4.8 Pembahasan Stabilitas Performa

Subbab ini menyatukan hasil pada Subbab 4.2 sampai dengan Subbab 4.7 untuk menjawab fokus penelitian, yaitu stabilitas performa model pada kondisi pengujian yang berbeda. Stabilitas dibaca melalui tiga hal: apakah peringkat model bertahan ketika kriteria

evaluasinya diganti, apakah galat model bertahan ketika periode ujinya berganti, dan apakah perbedaan yang terlihat cukup kuat untuk disebut signifikan.

Pemeriksaan pertama dilakukan dengan mengganti kriteria evaluasi. Kombinasi dataset dan model yang menempati posisi terbaik menurut delapan kriteria yang digunakan pada bab ini ditunjukkan pada Tabel 4.5.

Tabel 4.5 Kombinasi Terbaik menurut Kriteria Evaluasi

Kriteria evaluasi	**Kombinasi terbaik**	**Nilai**
Mean RMSE terendah	Random Forest Dataset B	0,044536
Mean MAE terendah	Random Forest Dataset B	0,033355
Directional accuracy tertinggi	Random Forest Dataset B	61,11%
Perbaikan terhadap *no-change forecast* tertinggi	Random Forest Dataset B	0,996%
Standar deviasi RMSE antar-*fold* terendah	XGBoost Dataset A	0,032274
Worst-fold RMSE terendah	XGBoost Dataset A	0,105896
RMSE rezim rendah terendah	Random Forest Dataset A	0,024915
RMSE rezim sedang terendah	Ridge Regression Dataset A	0,076941

Berdasarkan Tabel 4.5, tidak ada satu kombinasi pun yang unggul pada seluruh kriteria. Random Forest Dataset B unggul pada empat kriteria yang berkaitan dengan galat rata-rata dan ketepatan arah, tetapi tidak pada kriteria yang mengukur kestabilan galat antar-*fold* maupun pada perbandingan per rezim. Dari delapan kriteria yang digunakan, terdapat empat kombinasi berbeda yang pernah menempati posisi terbaik. Rezim tinggi tidak dimasukkan ke dalam tabel karena hanya memuat lima observasi dan berstatus eksploratif sesuai ketentuan pada Subbab 3.7.

Dua kriteria kestabilan pada tabel tersebut dimenangkan XGBoost Dataset A, yaitu kombinasi yang terdampak keterbatasan pada Subbab 4.2. Apabila kombinasi yang terdampak dikeluarkan, kriteria standar deviasi dimenangkan Ridge Regression Dataset A dan kriteria *Worst-fold* RMSE dimenangkan Random Forest Dataset A, sehingga jumlah kombinasi yang pernah menempati posisi terbaik menjadi tiga. Kesimpulan bahwa tidak ada kombinasi yang unggul pada seluruh kriteria tidak berubah pada kedua pembacaan tersebut.

Sumber variasi galat yang paling besar bukan pilihan model, melainkan periode pengujian. Rata-rata RMSE dua belas kombinasi bergerak dari 0,109953 pada 2020 menjadi 0,020044 pada 2022, sehingga selisih antartahun mencapai 0,089909 atau sekitar 5,5 kali lipat. Sebagai pembanding, selisih antara kombinasi terbaik dan terburuk di dalam satu tahun yang

sama memiliki nilai tengah 0,006563, dan pada 2025 hanya 0,000895. Selisih antartahun karena itu sekitar tiga belas kali lebih besar daripada selisih antarmodel di dalam satu tahun.

Urutan kesulitan antartahun juga hampir sama untuk semua kombinasi. Tahun 2020 menjadi tahun dengan galat tertinggi pada dua belas dari dua belas kombinasi, dan 2022 menjadi tahun dengan galat terendah pada sebelas dari dua belas kombinasi, dengan korelasi peringkat Spearman antarkombinasi rata-rata 0,9714 dan terendah 0,8286. Pola yang sama muncul pada metode pembanding sederhana. *No-change forecast* menghasilkan RMSE 0,108545 pada 2020 dan 0,020643 pada 2022 meskipun tidak menggunakan fitur maupun proses pelatihan sama sekali. Karena aturan prediksi yang tidak mempelajari apa pun mengikuti pola tahunan yang sama dengan seluruh model, kesulitan prediksi pada penelitian ini lebih tepat dibaca sebagai sifat periode datanya, bukan sebagai perbedaan kemampuan antarmodel.

Pembacaan yang sejalan muncul pada pengelompokan rezim volatilitas. Selisih rata-rata antara RMSE rezim sedang dan rezim rendah mencapai 0,052217, sedangkan sebaran antarkombinasi di dalam rezim rendah hanya 0,004380 dan di dalam rezim sedang 0,004750. Jarak antarrezim karena itu sekitar sebelas kali lebih besar daripada jarak antarmodel di dalam satu rezim. *No-change forecast* kembali mengikuti pola yang sama, dengan RMSE 0,025627 pada rezim rendah dan 0,078499 pada rezim sedang.

Hasil uji Diebold-Mariano pada Subbab 4.6 menjelaskan mengapa peringkat model tidak bertahan. Dari 120 perbandingan antarmodel pada *lag* utama, 28 perbandingan signifikan sebelum koreksi tetapi tidak satu pun bertahan setelah koreksi Holm, dengan nilai p terkoreksi terkecil 0,0767. Selisih *Mean* RMSE antarmodel di dalam satu dataset hanya berkisar antara 0,0014 dan 0,0059, sedangkan standar deviasi galat antar-*fold* berkisar antara 0,032274 dan 0,042761. Selisih yang jauh lebih kecil daripada variasinya sendiri memang sulit dibedakan dari variasi acak pada 72 observasi. Ketidakstabilan peringkat pada Tabel 4.5 karena itu konsisten dengan hasil pengujian statistik, bukan bertentangan dengannya.

Hasil ini tidak mendukung anggapan bahwa model yang lebih kompleks selalu menghasilkan prediksi yang lebih baik, maupun sebaliknya. Random Forest sebagai model berbasis pohon menempati *Mean* RMSE terendah pada ketiga dataset, sedangkan Ridge Regression sebagai model linear menempati posisi terbawah pada Dataset B dengan *Mean* RMSE tertinggi, yaitu 0,050440. Namun karena tidak ada perbandingan antarmodel yang bertahan setelah koreksi Holm, urutan tersebut tidak dapat dibaca sebagai keunggulan satu keluarga model atas keluarga lainnya.

Temuan tersebut sejalan dengan (Beck et al., 2025), yang menunjukkan bahwa model peramalan yang lebih kompleks tidak secara konsisten mengungguli *naive forecast* pada deret keuangan dengan tingkat prediktabilitas rendah. Kehati-hatian dalam menyimpulkan keunggulan model juga terlihat pada (Pagliaro, 2026), ketika bukti performa kerangka *regime-aware* LightGBM tidak mencapai signifikansi formal setelah memperhitungkan *multiple testing*. Meskipun objek, target prediksi, dan ukuran performanya berbeda, kedua penelitian tersebut mendukung pentingnya penggunaan metode pembanding sederhana dan pengujian statistik sebelum menyatakan satu model lebih unggul.

Penambahan kelompok fitur juga tidak mengubah kesimpulan tersebut. *Ablation study* pada Subbab 4.7 menunjukkan enam dari dua belas tahap menurunkan *Mean* RMSE dan enam lainnya menaikkannya, dengan arah perubahan yang berbeda antarmodel pada tahap yang sama. Satu-satunya perbedaan yang bertahan setelah koreksi Holm justru menunjukkan bahwa penambahan indikator pasar global memperburuk galat absolut Random Forest.

Berdasarkan seluruh pembacaan tersebut, penelitian ini tidak menemukan model yang dapat disebut stabil unggul. Yang stabil justru urutan kesulitannya, yaitu periode dan kondisi pasar tertentu menghasilkan galat yang lebih besar bagi seluruh model sekaligus, termasuk bagi aturan prediksi yang tidak dilatih sama sekali. Temuan ini perlu dibaca dengan memperhatikan cakupan evaluasi, yaitu 72 observasi bulanan pada enam *fold* tahunan, satu indeks, satu horizon prediksi, dan hanya lima observasi pada rezim volatilitas tinggi. Cakupan tersebut membatasi kekuatan uji dan menjadi salah satu alasan mengapa perbedaan kecil antarmodel sulit dibuktikan secara statistik. Implikasi dan saran atas keterbatasan ini disampaikan pada Bab 5.

4.9 Ringkasan Temuan Bab

Bab ini mengevaluasi empat model *machine learning* pada tiga dataset untuk memprediksi *return* IHSG 20 sesi perdagangan ke depan, menggunakan enam *fold* tahunan 2020 sampai dengan 2025 dengan 72 observasi bulanan pada tanggal pengujian yang identik untuk seluruh dataset.

Random Forest menghasilkan *Mean* RMSE terendah pada ketiga dataset, dengan nilai terendah 0,044536 pada Dataset B. Selisih antarmodel di dalam satu dataset berukuran kecil, yaitu antara 0,0014 dan 0,0059, sedangkan *coefficient of variation* seluruh kombinasi berada pada rentang 0,69 sampai 0,85.

Terhadap *no-change forecast*, hanya tiga dari dua belas kombinasi yang menghasilkan RMSE lebih rendah dan ketiganya adalah Random Forest, dengan perbaikan terbesar 0,996%

pada Dataset B. Terhadap *persistence forecast*, seluruh dua belas kombinasi lebih unggul dengan perbaikan antara 8,706% dan 19,392%. Nilai tambah model karena itu bergantung pada metode pembanding yang dipilih.

Evaluasi tahunan menunjukkan galat berubah tajam antarperiode. Seluruh model mencatat galat tertinggi pada 2020 dan hampir seluruhnya mencatat galat terendah pada 2022, dengan rata-rata RMSE 0,109953 dan 0,020044 pada kedua tahun tersebut.

Evaluasi berdasarkan rezim volatilitas menunjukkan galat tidak meningkat berurutan dari rezim rendah ke rezim tinggi. Galat terbesar terjadi pada rezim sedang pada sebelas dari dua belas kombinasi. Rezim tinggi hanya memuat lima observasi sehingga seluruh angkanya dibaca sebagai hasil eksploratif.

Uji Diebold-Mariano dengan koreksi Holm menunjukkan tidak satu pun dari 120 perbandingan antarmodel pada *lag* utama bertahan setelah koreksi, dengan nilai p terkoreksi terkecil 0,0767. Peringkat model pada bab ini karena itu bersifat deskriptif. *Ablation study* menunjukkan hanya satu dari dua belas tahap yang bertahan setelah koreksi, yaitu penambahan indikator pasar global yang memperburuk galat absolut Random Forest dengan nilai p terkoreksi 0,0213.

Secara keseluruhan, penelitian ini tidak menemukan model yang stabil unggul pada seluruh kriteria maupun seluruh kondisi pengujian. Variasi galat yang berasal dari periode pengujian dan kondisi pasar jauh lebih besar daripada variasi yang berasal dari pilihan model, dan pola tersebut juga berlaku pada metode pembanding yang tidak dilatih. Rangkuman temuan ini menjadi dasar penyusunan kesimpulan dan saran pada Bab 5.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil eksperimen dan pembahasan pada Bab 4, penelitian ini memperoleh beberapa kesimpulan utama. Kesimpulan berikut disusun sesuai dengan tujuan penelitian, yaitu mengevaluasi performa model *machine learning* dalam memprediksi *return* IHSG 20 sesi perdagangan ke depan pada rezim volatilitas yang berbeda. Seluruh kesimpulan berlaku pada cakupan evaluasi penelitian ini, yaitu enam *fold* tahunan 2020 sampai dengan 2025 dengan 72 observasi bulanan untuk setiap kombinasi dataset dan model.

1. Random Forest menghasilkan galat rata-rata paling rendah dalam eksperimen ini. Berdasarkan *Mean* RMSE, Random Forest menempati posisi terendah pada Dataset A, Dataset B, dan Dataset C, dengan nilai terendah 0,044536 pada Dataset B. Namun uji Diebold-Mariano dengan koreksi Holm menunjukkan bahwa tidak satu pun dari 120 perbandingan antarmodel pada *lag* utama bertahan setelah koreksi, dengan nilai p terkoreksi terkecil 0,0767. Random Forest karena itu hanya dapat disebut terbaik secara deskriptif pada sampel evaluasi ini dan belum terbukti lebih akurat secara statistik daripada Ridge Regression, LightGBM, maupun XGBoost.
2. Perbedaan antara Dataset A, Dataset B, dan Dataset C terhadap galat prediksi berukuran kecil dan arahnya tidak seragam antarmodel. Pada Random Forest, selisih *Mean* RMSE antara Dataset B dan Dataset A hanya 0,000273 dan antara Dataset B dan Dataset C hanya 0,000189. *Ablation study* menunjukkan enam dari dua belas tahap penambahan kelompok fitur menurunkan *Mean* RMSE dan enam lainnya menaikkannya. Hanya satu perbedaan yang bertahan setelah koreksi Holm, yaitu penambahan indikator pasar global pada Random Forest yang justru memperburuk galat absolut dengan nilai p terkoreksi 0,0213. Penambahan informasi makroekonomi dan volatilitas karena itu belum terbukti memberikan kontribusi prediktif yang konsisten.
3. Nilai tambah model terhadap metode pembanding sederhana bergantung pada pembanding yang dipilih. Terhadap *persistence forecast*, seluruh dua belas kombinasi menghasilkan RMSE lebih rendah dengan perbaikan antara 8,706% dan 19,392%. Terhadap *no-change forecast*, hanya tiga dari dua belas kombinasi yang lebih rendah dan ketiganya adalah Random Forest, dengan perbaikan terbesar hanya 0,996% pada

Dataset B. Tidak satu pun perbandingan terhadap kedua metode pembanding tersebut bertahan setelah koreksi Holm. Dengan demikian, model *machine learning* belum terbukti memberikan nilai tambah yang kuat terhadap aturan prediksi yang selalu memprediksi nol.

4. Galat prediksi berubah tajam antarperiode uji. Rata-rata RMSE dua belas kombinasi bergerak dari 0,109953 pada 2020 menjadi 0,020044 pada 2022, atau sekitar 5,5 kali lipat. Seluruh kombinasi mencatat galat tertinggi pada 2020 dan hampir seluruhnya mencatat galat terendah pada 2022. Temuan ini menunjukkan bahwa nilai rata-rata RMSE tidak cukup untuk menilai perilaku model karena rata-rata dapat menyembunyikan periode ketika galat meningkat tajam.
5. Galat prediksi tidak meningkat secara berurutan dari rezim volatilitas rendah ke rezim tinggi. Galat terbesar justru terjadi pada rezim sedang pada sebelas dari dua belas kombinasi. Rezim tinggi hanya memuat lima observasi dari 72 observasi pengujian sehingga seluruh angkanya dibaca sebagai hasil eksploratif dan tidak digunakan sebagai dasar kesimpulan stabilitas antarrezim.

Secara keseluruhan, penelitian ini tidak menemukan model yang dapat disebut stabil unggul. Peringkat model berubah ketika kriteria evaluasinya diganti, dan variasi galat yang berasal dari periode pengujian jauh lebih besar daripada variasi yang berasal dari pilihan model. Selisih rata-rata RMSE antartahun mencapai 0,089909, sedangkan selisih antarkombinasi di dalam satu tahun memiliki nilai tengah 0,006563. Pola tahunan yang sama juga terjadi pada *no-change forecast* yang tidak menggunakan fitur maupun proses pelatihan. Kesulitan prediksi *return* IHSG 20 sesi perdagangan ke depan karena itu lebih tepat dibaca sebagai sifat periode datanya daripada sebagai perbedaan kemampuan antarmodel.

5.2 Saran

Berdasarkan hasil dan keterbatasan penelitian, terdapat beberapa saran yang dapat dipertimbangkan untuk penelitian selanjutnya. Keterbatasan utama penelitian ini adalah cakupan evaluasi yang terbatas, yaitu 72 observasi bulanan pada enam *fold* tahunan, satu indeks, satu horizon prediksi, dan hanya lima observasi pada rezim volatilitas tinggi. Saran berikut diarahkan pada pengembangan penelitian, bukan pada penggunaan model sebagai dasar keputusan investasi langsung.

1. Penelitian selanjutnya dapat memperbesar jumlah observasi evaluasi agar pengujian statistik memiliki daya uji yang lebih memadai. Pada penelitian ini seluruh

perbandingan antarmodel gugur setelah koreksi Holm karena selisih galat antarmodel jauh lebih kecil daripada variasi galat antar-*fold* pada 72 observasi yang tersedia. Penambahan observasi dapat dilakukan melalui periode pengujian yang lebih panjang atau frekuensi *anchor* yang lebih rapat dengan tetap memperhatikan *overlap* target. Cara tersebut sekaligus dapat menambah jumlah observasi pada rezim volatilitas tinggi sehingga rezim tersebut dapat dibandingkan secara formal, bukan hanya secara eksploratif.

2. Penelitian selanjutnya perlu memeriksa ruang pencarian *hyperparameter* agar sesuai dengan skala target yang diprediksi. Pada penelitian ini LightGBM pada *fold* 2020 dan XGBoost pada *fold* 2022 menghasilkan prediksi konstan karena proses penalaan memilih nilai regularisasi yang sangat besar, sehingga model tidak menggunakan fitur pada *fold* tersebut. Pembatasan rentang regularisasi atau penolakan konfigurasi yang menghasilkan model tanpa percabangan dapat mencegah kondisi tersebut terulang.
3. Penelitian selanjutnya dapat menguji horizon prediksi yang berbeda, misalnya *return* 5 sesi, 10 sesi, atau 21 sesi perdagangan. Penelitian ini hanya berfokus pada *return* 20 sesi perdagangan ke depan, sehingga hasilnya tidak dapat langsung digeneralisasi untuk horizon harian, mingguan, atau horizon lain. Pengujian beberapa horizon dapat membantu melihat apakah performa model berubah ketika target prediksi dibuat lebih pendek atau lebih panjang.
4. Penelitian berikutnya dapat menambahkan variabel eksternal lain yang berpotensi berkaitan dengan pergerakan IHSG. Variabel tersebut dapat mencakup sentimen berita, indikator fundamental, indeks sektoral, atau indikator pasar global lain. Namun, penambahan variabel perlu dilakukan secara hati-hati karena hasil penelitian ini menunjukkan bahwa tambahan fitur tidak selalu menurunkan galat prediksi.
5. Penelitian lanjutan dapat menguji model lain atau pendekatan *ensemble* yang berbeda, tetapi tetap perlu dibandingkan dengan metode pembanding sederhana seperti *no-change forecast* dan *persistence forecast*. Perbandingan dengan metode pembanding sederhana penting agar model yang lebih kompleks tidak hanya terlihat lebih baik dibandingkan model lain, tetapi juga memberikan nilai tambah dibandingkan aturan prediksi sederhana.
6. Evaluasi dapat diperluas ke ukuran berbasis manfaat ekonomi jika penelitian diarahkan pada penggunaan prediksi untuk strategi investasi. Contoh ukuran yang dapat dipertimbangkan adalah *Sharpe ratio* dan *maximum drawdown*. Namun, ukuran

tersebut sebaiknya digunakan apabila penelitian memang menguji strategi keputusan investasi, bukan hanya mengevaluasi galat prediksi *return*.

7. Penelitian selanjutnya dapat melakukan pengujian pada periode data yang berbeda atau memperbarui data ketika observasi baru tersedia. Hal ini penting karena hasil penelitian menunjukkan bahwa performa model berubah dari tahun ke tahun. Pengujian pada periode tambahan dapat membantu menilai apakah pola hasil yang ditemukan dalam penelitian ini tetap bertahan pada data baru.

DAFTAR PUSTAKA

Adinda, N. P., Salamy, F. I., Subagyo, S. F., Sari, D. P., & Zuhdi, S. (2026). Forecasting the IHSG with an ARIMA-ARCH Model to Address Heteroscedasticity Effects. *Bit-Tech (Binary Digital - Technology)*, *8*(3). https://doi.org/10.32877/bt.v8i3.3515

Beck, N., Dovern, J., & Vogl, S. (2025). Mind the Naive Forecast! A Rigorous Evaluation of Forecasting Models for Time Series with Low Predictability. *Applied Intelligence*, *55*, 395. https://doi.org/10.1007/s10489-025-06268-w

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327. https://doi.org/10.1016/0304-4076(86)90063-1

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Campbell, J. Y., & Thompson, S. B. (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies*, *21*(4), 1509–1531. https://doi.org/10.1093/rfs/hhm055

Cao, Y. (2025). Reevaluating Stock Price Prediction: The Influence of Machine Learning on Forecast Accuracy. *Proceedings of the 1st International Conference on E-Commerce and Artificial Intelligence (ECAI 2024)*, 85–91. https://doi.org/10.5220/0013207000004568

Chen, T., & Guestrin, C. (2016). XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Deep, A., Monico, C., Shirvani, A., Rachev, S., & Fabozzi, F. J. (2024). *Assessing the Impact of Technical Indicators on Machine Learning Models for Stock Price Prediction*. arXiv. https://doi.org/10.48550/arXiv.2412.15448

Deep, G., Deep, A., & Lamptey, W. (2025). *Interpretable Hypothesis-Driven Trading: A Rigorous Walk-Forward Validation Framework for Market Microstructure Signals*. Retrieved from https://arxiv.org/abs/2512.12924

Diebold, F. X., & Mariano, R. S. (2002). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144. https://doi.org/10.1198/073500102753410444

Engle, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica*, *50*(4), 987. https://doi.org/10.2307/1912773

Fama, E. F. (1995). Random Walks in Stock-Market Prices. *Financial Analysts Journal*, *51*, 75–80. https://doi.org/10.2469/faj.v51.n1.1861

Galena, M. V., Sediono, Mardianto, M. F. F., & Pusporani, E. (2025). Prediction Analysis of Jakarta Composite Index Movement Using Support Vector Regression Method. *G-Tech: Jurnal Teknologi Terapan*, *9*(1), 68–77. https://doi.org/10.70609/gtech.v9i1.5879

Gomez Grau, C. (2025). *Bayesian-Optimized Machine Learning Models for Stock Return Prediction: A Regime-Based Approach*.

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, *57*(2), 357. https://doi.org/10.2307/1912559

Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices. *Data Mining and Knowledge Discovery*, *37*, 788–832. https://doi.org/10.1007/s10618-022-00894-5

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Indra, Supian, S., Sukono, Riaman, Saputra, M. P. A., Azahra, A. S., & Pirdaus, D. I. (2025). Stock Return Prediction on the LQ45 Market Index in the Indonesia Stock Exchange Using a Machine Learning Algorithm Based on Technical Indicators. *Journal of Risk and Financial Management*, *18*(12), 714. https://doi.org/10.3390/jrfm18120714

Ismanto, H. (2026). Forecasting the Jakarta Composite Index (IHSG) Using the Moving Average Method. *JSIKTI (Jurnal Sistem Informasi Dan Komputer Terapan Indonesia)*, *8*(3), 45–53. https://doi.org/10.33173/jsikti.283

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.

Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003

Kolassa, S. (2026). The Symmetric Mean Absolute Percentage Error: Unnecessary or Dangerous. *Forecasting*, *8*(2), 24. https://doi.org/10.3390/forecast8020024

Liu, T. (2026). A Comparative Study of Transformer-Based and Classical Models for Financial Time-Series Forecasting. *Journal of Risk and Financial Management*, *19*(3), 203. https://doi.org/10.3390/jrfm19030203

Lutey, M. (2022). Robust Testing for Bollinger Band, Moving Average and Relative Strength Index. *Journal of Finance Issues*, *20*(1), 27–46. https://doi.org/10.58886/jfi.v20i1.3218

Marpaung, N. N., & Pangestuti, I. R. D. (2024). Macroeconomic Factors and Jakarta Stock Exchange: A Comparative Analysis Pre- and Until the COVID-19 Pandemic. *Sage Open*, *14*(2). https://doi.org/10.1177/21582440241247894

Muhammad, D. (2024). An Explainable Deep Learning Approach for Stock Market Trend Prediction. *Heliyon*, *10*, e40095. https://doi.org/10.1016/j.heliyon.2024.e40095

Najibullah, Azlina, Ariansyah, R., & Lisnawati. (2024). Prediksi Indeks Harga Saham Gabungan (IHSG) Menggunakan Regresi K-Nearest Neighbors. *Jurnal Kreasi Rakyat*, *2*(1), 15–23.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703. https://doi.org/10.2307/1913610

Pagliaro, A. (2026). Regime-Aware LightGBM for Stock Market Forecasting: A Validated Walk-Forward Framework with Statistical Rigor and Explainable AI Analysis. *Electronics*. https://doi.org/10.3390/electronics15061334

Saputra, M. P. A., Setyawan, D. P., & Wahid, A. J. (2025). Indonesian Banking Stock Portfolio Optimization Based on Ridge Regression Prediction. *International Journal of Business, Economics and Social Development*, *6*(2), 330–337. https://doi.org/10.46336/ijbesd.v6i2.1064

Situmorang, T. V., & Kaluge, D. (2025). The Effects of Macroeconomic Factors and Singapore and Thailand Stock Indices on The Volatility of the Jakarta Composite Index. *Contemporary Studies in Economic, Finance and Banking*, *4*(4), 903–916. https://doi.org/10.21776/csefb.2025.04.4.12

Spulbăr, C., & Ene, C. C. (2025). Rolling Walk-Forward LSTM for Daily Stock-Return Prediction in European Defence Markets. *Studies in Business and Economics*, *20*(3), 240–260. https://doi.org/10.2478/sbe-2025-0053

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, *16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Tsang, W. W. H., & Chong, T. T. L. (2009). Profitability of the On-Balance Volume Indicator. *Economics Bulletin*, *29*(3), 2424–2431.

Wahyuddin, E. P., Caraka, R. E., Kurniawan, R., Caesarendra, W., Gio, P. U., & Pardamean, B. (2025). Improved LSTM hyperparameters alongside sentiment walk-forward validation for time series prediction. *Journal of Open Innovation: Technology, Market, and Complexity*, *11*(1), 100458. https://doi.org/10.1016/j.joitmc.2024.100458

Wiratama, R. K. P., Saikhu, A., & Suciati, N. (2025). GARCH Model Hybridization with Feed Forward Neural Network Algorithm Approach for Predicting the Volatility of the Composite Stock Price Index. *JUTI: Jurnal Ilmiah Teknologi Informasi*, *23*(2), 88–110. https://doi.org/10.12962/j24068535.v23i1.a1278

LAMPIRAN