



Komparasi Akurasi Model SARIMA, LSTM, dan GRU dalam Peramalan Indeks Harga Saham S&P 500

Imam Adiyana^{1,*}, Angga Kurniawan², Bella Okta Sari Miranda¹

¹ Fakultas Ekonomi Bisnis, Program Studi S1 Bisnis Digital, Universitas Telkom, Purwokerto, Indonesia

² Fakultas Informatika, Program Studi S1 Sains Data, Universitas Telkom, Purwokerto, Indonesia

Email: ^{1,*}imamadiyana@telkomuniversity.ac.id, ²anggakurniawan@telkomuniversity.ac.id,

³bellaoktasarimiranda@telkomuniversity.ac.id,

Email Penulis Korespondensi: imamadiyana@telkomuniversity.ac.id

Abstrak—Pergerakan indeks harga saham S&P 500 bersifat non-linear dan dipengaruhi oleh banyak faktor makroekonomi, sehingga peramalan harga penutupan (closing price) menjadi tantangan penting bagi investor dan manajer portofolio dalam pengambilan keputusan investasi. Penelitian ini bertujuan untuk mengetahui apakah model deep learning lebih baik dibandingkan model statistika klasik dalam meramalkan data non-linear pada data harga penutupan harian indeks S&P 500 dari dataset publik. Model peramalan deret waktu yang digunakan adalah Seasonal Autoregressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM) dan Gated Recurrent Unit (GRU). Ketiga model dibangun dan dilatih secara independen, kemudian dievaluasi menggunakan Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), dan koefisien determinasi (R-squared). Hasil penelitian menunjukkan bahwa model GRU memberikan performa terbaik dengan MAE sebesar 87,04, RMSE sebesar 100,72, MAPE sebesar 2,94 persen, dan R-squared sebesar 0,836, diikuti oleh model LSTM dengan MAE 105,89, RMSE 132,13, MAPE 3,55 persen, dan R-squared 0,718. Model SARIMA menunjukkan performa jauh lebih rendah dengan MAE 418,77, RMSE 610,20, MAPE 13,72 persen, dan R-squared negatif sebesar -5,02, yang mengindikasikan model tersebut tidak mampu menangkap pola non-linear dan tren kuat pada data saham. Temuan ini menegaskan bahwa model deep learning, khususnya GRU, lebih unggul dibandingkan model statistik klasik dalam meramalkan indeks saham dengan volatilitas tinggi, sekaligus menawarkan efisiensi komputasi yang lebih baik dibandingkan LSTM.

Kata Kunci: GRU; LSTM; SARIMA; Indeks S&P 500; Peramalan Harga Saham

Abstract—Movement of S&P 500 stock price index is non-linear and influenced by numerous macroeconomic factors, making closing price forecasting a critical challenge for investors and portfolio managers in investment decision-making. This study aims to determine whether deep learning models outperform classical statistical models in forecasting non-linear patterns in the daily closing price data of the S&P 500 index using a public dataset. The forecasting models employed are Seasonal Autoregressive Integrated Moving Average (SARIMA), Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). The three models were independently constructed and trained, then evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R-squared). Results show that GRU delivered the best performance, with an MAE of 87.04, RMSE of 100.72, MAPE of 2.94 percent, and R-squared of 0.836, followed by LSTM with an MAE of 105.89, RMSE of 132.13, MAPE of 3.55 percent, and R-squared of 0.718. SARIMA performed considerably worse, with an MAE of 418.77, RMSE of 610.20, MAPE of 13.72 percent, and a negative R-squared of -5.02, indicating its inability to capture the non-linear patterns and strong trends in the data. These findings confirm that deep learning models, particularly GRU, outperform classical statistical models in forecasting highly volatile stock indices, while also offering better computational efficiency than LSTM.

Keywords: GRU; LSTM; SARIMA; S&P 500 Index; Stock Price Forecasting

1. PENDAHULUAN

Indeks S&P 500 merupakan salah satu indeks pasar saham paling berpengaruh di dunia yang merepresentasikan kinerja 500 perusahaan publik terbesar di Amerika Serikat berdasarkan kapitalisasi pasar dan likuiditas. Indeks ini mencakup sekitar 80% dari total kapitalisasi pasar ekuitas AS dan menjadi acuan utama bagi investor global, manajer portofolio, serta analis keuangan dalam menilai kondisi pasar modal dan kesehatan ekonomi dunia [1]. Sepanjang dekade terakhir, indeks S&P 500 telah menunjukkan pertumbuhan yang sangat pesat, ditutup di kisaran 4.700-6.000 poin pada tahun 2024, didorong oleh performa solid perusahaan-perusahaan teknologi besar, serta ekspektasi penurunan suku bunga acuan *The Federal Reserve* [2]. Namun demikian, pergerakan harian indeks S&P 500 pada dasarnya bersifat *non-linear* dan *noisy*, serta dipengaruhi oleh rilis laporan keuangan emiten, dinamika makroekonomi, ketidakpastian geopolitik, ekspektasi suku bunga, dan bias perilaku investor [3][4]. Karakteristik deret waktu tersebut membuat prediksi harga penutupan (*closing price*) indeks S&P 500 menjadi tantangan yang signifikan. Studi terkini menunjukkan bahwa volatilitas pasar dapat menurunkan akurasi model prediksi harga saham [5]. Pada penelitian ini, data yang digunakan adalah data historis harian indeks S&P 500 yang bersumber dari dataset publik di Kaggle, yaitu "S&P 500 stock data" karya Cam Nugent, yang menyediakan data OHLCV (*Open, High, Low, Close, Volume*) harian untuk seluruh konstituen indeks S&P 500 [6].

Berbagai pendekatan telah diusulkan untuk memprediksi harga saham dan indeks pasar modal, mulai dari model statistik klasik hingga model *deep learning*. Model *Seasonal Autoregressive Integrated Moving Average* (SARIMA) dan variannya digunakan untuk menangkap pola linear dan musiman pada deret waktu, sedangkan pendekatan hibrida terkini menggabungkan kekuatan model statistik dan *deep learning* untuk meningkatkan kemampuan peramalan pada data keuangan [7][8]. Namun demikian, SARIMA dibangun di atas asumsi linearitas, stasioneritas, dan normalitas residual, sehingga kurang optimal ketika diterapkan pada data indeks saham yang pada umumnya mengandung komponen *non-linear* yang kuat, *regime switching*, serta volatilitas yang bervariasi seiring waktu [9]. Di sisi lain, arsitektur jaringan saraf berulang seperti *Long Short-Term Memory* (LSTM) dan *Gated Recurrent Unit* (GRU) mampu mempelajari dependensi



temporal dan pola *non-linear* yang kompleks pada data sekuensial [10][11]. LSTM menggunakan tiga gerbang (*input gate*, *forget gate*, dan *output gate*) ditambah *cell state* untuk mengontrol aliran informasi, sehingga efektif menangani dependensi jangka panjang [10]. Sementara itu, GRU menggunakan dua gerbang utama, yaitu *reset gate* dan *update gate*, dengan struktur yang lebih sederhana dan jumlah parameter yang umumnya lebih sedikit dibandingkan LSTM. Studi terkini menunjukkan bahwa GRU dapat menghasilkan akurasi yang lebih baik pada data saham tertentu, meskipun waktu pelatihan tetap bergantung pada konfigurasi model dan lingkungan komputasi [12][13]. Ketiga model tersebut, yaitu SARIMA, LSTM, dan GRU, masing-masing memiliki kekuatan dan keterbatasan yang saling melengkapi. SARIMA unggul dalam menangkap pola linear dan musiman eksplisit, LSTM menawarkan representasi jangka panjang yang kaya, dan GRU menyediakan alternatif yang lebih ringkas. Oleh karena itu, studi komparasi yang membandingkan ketiganya secara langsung pada data historis indeks S&P 500 menjadi langkah penting untuk menentukan model mana yang paling sesuai untuk peramalan indeks pasar modal global.

Sejumlah penelitian terdahulu telah membandingkan model statistik dengan model *deep learning* untuk peramalan deret waktu keuangan. Kirelli [14] membandingkan ARIMA dan LSTM pada harga saham perusahaan teknologi dan menunjukkan bahwa LSTM menghasilkan nilai kesalahan yang lebih rendah pada beberapa metrik, sedangkan ARIMA tetap kompetitif pada ukuran akurasi proporsional. Şengül Abedi [15] membandingkan SARIMA, LSTM, dan GRU pada enam deret waktu perdagangan luar negeri sektor pertanian dan pangan. Hasil penelitian tersebut menunjukkan bahwa tidak ada satu model yang selalu unggul, karena SARIMA lebih kuat pada seri dengan pola musiman yang jelas, sedangkan GRU lebih baik pada beberapa seri yang lebih *non-linear*. Hossain *et al.* [16] membandingkan empat pendekatan, yaitu LSTM, GRU, CNN, dan ARIMA, untuk memprediksi harga penutupan harian sektor keuangan S&P 500 dan menemukan bahwa LSTM mengungguli pendekatan lainnya pada sebagian besar skenario *train-test split* dengan perbaikan RMSE lebih dari 15% dan MAE lebih dari 30% dibanding ARIMA. Berradi *et al.* [17] secara khusus membandingkan LSTM dan GRU pada data harian S&P 500 periode 1991-2021 dan melaporkan bahwa GRU memberikan MSE dan MAE yang sedikit lebih rendah dibanding LSTM. Li dan Chen [18] membandingkan LSTM, GRU, dan *hybrid* LSTM-GRU untuk prediksi indeks S&P 500 dan menemukan bahwa LSTM tunggal memberikan performa terbaik berdasarkan RMSE, MSE, dan MAE. Yunita *et al.* [19] mengevaluasi RNN, LSTM, GRU, dan beberapa arsitektur hibrida melalui pengujian berulang. Hasilnya menunjukkan bahwa tidak ada arsitektur yang optimal secara universal dan performa model bergantung pada karakteristik pola temporal dari data yang dianalisis.

Berdasarkan tinjauan terhadap penelitian-penelitian terdahulu, dapat diidentifikasi beberapa celah penelitian sebagai berikut: (1) mayoritas penelitian komparasi model *time series* pada indeks S&P 500 hanya membandingkan dua model sekaligus, misalnya LSTM vs GRU [17][18] atau LSTM vs ARIMA [16]. Studi tiga model terbaru yang membandingkan SARIMA, LSTM, dan GRU lebih banyak diterapkan pada domain ekonomi lain [15], sehingga perbandingan ketiga model secara simultan pada data historis indeks S&P 500 masih terbatas; (2) sebagian besar penelitian terdahulu pada indeks S&P 500 berfokus pada perbandingan akurasi prediksi tanpa sebagian besar penelitian terdahulu pada indeks S&P 500 berfokus pada perbandingan akurasi prediksi tanpa melakukan analisis mendalam terhadap kekuatan dan kelemahan masing-masing model dalam menangkap karakteristik spesifik indeks, seperti efek *day-of-week*, *volatility clustering*, atau *regime switching* pada periode krisis keuangan [9]; serta (3) penelitian yang secara eksplisit mengevaluasi *trade-off* antara akurasi prediksi, kompleksitas parameter, dan efisiensi komputasi, khususnya antara LSTM dan GRU yang secara arsitektural serupa, pada konteks prediksi indeks S&P 500 masih relatif jarang dilakukan [13]. Celah-celah inilah yang menjadi dasar dan posisi kebaruan (*novelty*) dari penelitian ini, yaitu melakukan studi komparasi komprehensif antara tiga model, yaitu SARIMA, LSTM, dan GRU, pada data historis harian indeks S&P 500 yang bersumber dari dataset publik Kaggle.

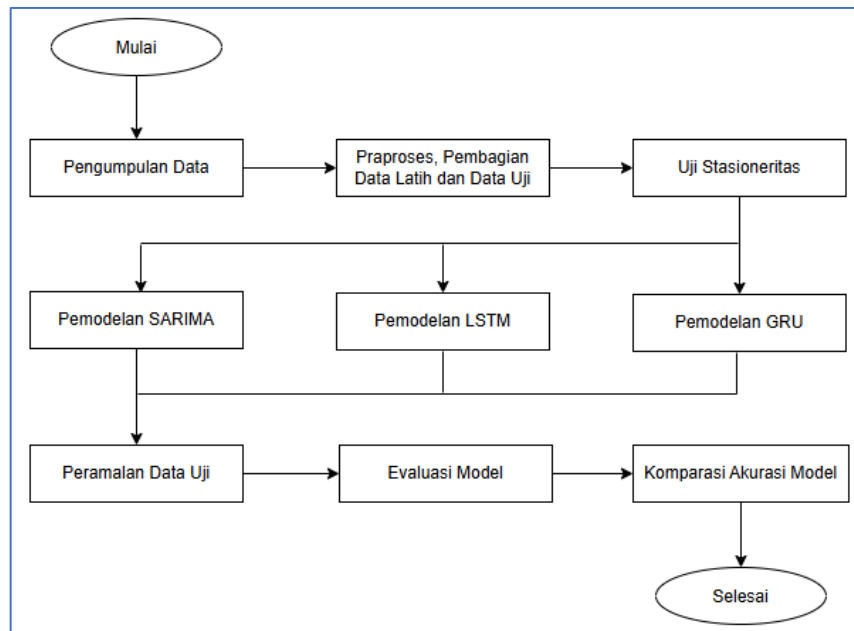
Berdasarkan uraian permasalahan di atas, penelitian ini bertujuan untuk: (1) merancang dan mengimplementasikan tiga model peramalan, yaitu SARIMA, LSTM, dan GRU, untuk memprediksi harga penutupan harian (*closing price*) indeks S&P 500 menggunakan data historis dari dataset Kaggle; (2) membandingkan performa ketiga model secara kuantitatif menggunakan metrik evaluasi *Mean Absolute Error* (MAE), *Root Mean Square Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE); serta (3) menganalisis kekuatan dan kelemahan masing-masing model dalam menangkap karakteristik spesifik data indeks S&P 500, khususnya pola tren, musiman, dan komponen *non-linear*, sebagai dasar rekomendasi pemilihan model yang paling sesuai untuk peramalan indeks pasar modal global. Penelitian ini dapat memberikan kontribusi baik secara akademis maupun praktis. Secara akademis, penelitian ini memperkaya khazanah literatur mengenai studi komparasi model *time series* pada domain prediksi indeks S&P 500, khususnya konfigurasi tiga model yang berbeda paradigma, yaitu model statistik (SARIMA) versus model *deep learning* (LSTM dan GRU), yang menggunakan dataset publik Kaggle sebagai sumber data. Secara praktis, hasil perbandingan model dapat menjadi rujukan bagi investor, analis, dan manajer portofolio global dalam memilih model prediksi yang paling akurat dan efisien secara komputasi untuk mendukung pengambilan keputusan investasi, manajemen portofolio, serta pengendalian risiko pada indeks pasar modal, khususnya S&P 500.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menerapkan pendekatan komparasi antara satu model statistik klasik, yaitu *Seasonal Autoregressive Integrated Moving Average* (SARIMA), dengan dua model *deep learning* berbasis *recurrent neural network*, yaitu *Long*

Short-Term Memory (LSTM) dan *Gated Recurrent Unit* (GRU), untuk meramalkan pergerakan indeks harga saham S&P 500. Ketiga model dibangun dan dilatih secara independen di atas data historis yang sama, kemudian akurasi peramalannya dibandingkan menggunakan metrik evaluasi yang identik agar diperoleh perbandingan yang adil, sejalan dengan pendekatan yang digunakan pada beberapa penelitian komparasi model peramalan deret waktu sebelumnya [15][21][25][26]. Kerangka penelitian disusun mengikuti tahapan sistematis mulai dari pengumpulan data hingga evaluasi model. Tahapan penelitian secara lengkap digambarkan pada diagram alir Gambar 1.



Gambar 1. Tahapan Penelitian

Tahap pertama adalah pengumpulan data, yaitu data historis harga penutupan (*closing price*) harian indeks saham S&P 500 yang diperoleh dari sumber data pasar keuangan. Selanjutnya dilakukan praproses data, meliputi penanganan nilai yang hilang akibat hari libur bursa (*non-trading days*), serta pembagian data menjadi data latih dan data uji. Untuk kebutuhan pemodelan LSTM dan GRU, data dinormalisasi menggunakan *Min-Max Scaling* ke rentang (0,1), sedangkan model SARIMA menggunakan data pada skala aslinya. Tahap berikutnya adalah uji stasioneritas menggunakan *Augmented Dickey-Fuller Test* (ADF Test) untuk menentukan perlunya proses *differencing* sebelum pemodelan SARIMA dilakukan [20]. Pada tahap pemodelan SARIMA, parameter model order (p,d,q) dan *seasonal* order (P,D,Q,s) ditentukan secara otomatis menggunakan algoritma *stepwise Hyndman-Khandakar* yang diimplementasikan pada pustaka *pmdarima* (fungsi *auto_arima*), berdasarkan nilai *Akaike Information Criterion* (AIC) terendah [8]. Model SARIMA yang diperoleh selanjutnya digunakan untuk menghasilkan prediksi pada data uji sebagai representasi pendekatan statistik klasik dalam peramalan deret waktu keuangan.

Pada tahap pemodelan LSTM, data harga saham yang telah dinormalisasi disusun ke dalam bentuk *sliding window* berupa rangkaian nilai historis (*lag*) sebagai input untuk memprediksi nilai pada periode berikutnya. Arsitektur LSTM terdiri atas satu atau beberapa lapisan LSTM yang dilengkapi *forget gate*, *input gate*, dan *output gate* untuk mengatur aliran informasi jangka panjang, dilanjutkan dengan lapisan *dropout* guna mencegah *overfitting* dan lapisan *dense* sebagai keluaran akhir [10][24]. Model dilatih menggunakan *optimizer* Adam dengan fungsi kerugian *Mean Squared Error* (MSE). Tahap pemodelan GRU, struktur input berupa *sliding window* yang sama diterapkan, namun unit rekuren yang digunakan adalah GRU yang memiliki dua gerbang utama, yaitu *update gate* dan *reset gate*. Dibandingkan LSTM, GRU memiliki struktur gerbang yang lebih sederhana karena tidak memisahkan *cell state* dan *hidden state*, sehingga secara komputasi lebih efisien namun tetap mampu menangkap dependensi temporal dan pola *non-linear* pada data deret waktu [11][24]. Model GRU dilatih dengan konfigurasi *optimizer* dan fungsi kerugian yang sama dengan model LSTM agar perbandingan kinerja antar model seimbang. Tahap terakhir adalah evaluasi model, yaitu membandingkan akurasi peramalan ketiga model, SARIMA, LSTM, dan GRU pada data uji yang sama menggunakan tiga metrik, yaitu *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), dan *Mean Absolute Percentage Error* (MAPE) [8][22]. Model dengan nilai ketiga metrik terendah dianggap memiliki akurasi peramalan terbaik terhadap pergerakan indeks harga saham S&P 500.

2.2 Formulasi Model SARIMA, LSTM, dan GRU

Model SARIMA secara umum dinotasikan sebagai $SARIMA(p,d,q)(P,D,Q)s$, yang merupakan gabungan antara komponen *non-seasonal* $ARIMA(p,d,q)$ dan komponen musiman $(P,D,Q)s$, dengan bentuk umum sebagaimana persamaan (1) [8]:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D y_t = \theta() \Theta(B^s) \varepsilon_t \quad (1)$$



dengan y_t adalah nilai indeks harga saham S&P 500 pada waktu t , B adalah *backshift operator*, $\phi(B)$ dan $\theta(B)$ masing-masing adalah polinomial AR dan MA non-musiman, $\Phi(Bs)$ dan $\Theta(Bs)$ adalah polinomial AR dan MA musiman, serta ϵ_t adalah *white noise* [8]. Model LSTM memproses data deret waktu x_t melalui unit sel memori yang diatur oleh tiga gerbang utama, yaitu *forget gate* (f_t), *input gate* (i_t), dan *output gate* (o_t), sebagaimana dirumuskan pada persamaan (2) sampai (6) [10][24]:

$$f_t = \sigma(W^f \cdot [h_{t-1}, x_t] + b^f) \quad (2)$$

$$i_t = \sigma(W^i \cdot [h_{t-1}, x_t] + b^i) \quad (3)$$

$$\tilde{C}_t = \tanh(W^c \cdot [h_{t-1}, x_t] + b^c) \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W^o \cdot [h_{t-1}, x_t] + b^o), h_t = o_t \odot \tanh(C_t) \quad (6)$$

dengan σ adalah fungsi aktivasi *sigmoid*, C_t adalah *cell state* pada waktu t , W dan b adalah bobot dan *bias* yang dipelajari selama pelatihan, serta h_t adalah *hidden state* pada waktu t [10]. Model GRU menyederhanakan mekanisme gerbang pada LSTM dengan menggabungkan *forget gate* dan *input gate* menjadi satu *update gate* (z_t), serta menambahkan *reset gate* (r_t), sebagaimana dirumuskan pada persamaan (7) sampai (9) [11][24]:

$$z_t = \sigma(W^z \cdot [h_{t-1}, x_t] + b^z) \quad (7)$$

$$r_t = \sigma(W^r \cdot [h_{t-1}, x_t] + b^r) \quad (8)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t \odot h_{t-1}, x_t] + b), h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (9)$$

dengan z_t mengatur proporsi informasi masa lalu yang dipertahankan, r_t menentukan seberapa besar *hidden state* sebelumnya diabaikan dalam menghitung kandidat *hidden state* $\tilde{h}_t$, dan h_t adalah *hidden state* pada waktu t [11]. Kinerja ketiga model dievaluasi menggunakan tiga metrik, yaitu MAE, RMSE, dan MAPE, yang masing-masing dirumuskan pada persamaan (10) hingga (12) [8][22]:

$$MAE = (1/n) \sum |y_i - \hat{y}_i| \quad (10)$$

$$RMSE = \sqrt{(1/n) \sum (y_i - \hat{y}_i)^2} \quad (11)$$

$$MAPE = (100\%/n) \sum |(y_i - \hat{y}_i)/y_i| \quad (12)$$

Nilai MAE dan RMSE merepresentasikan besarnya kesalahan prediksi dalam satuan poin indeks, sementara MAPE merepresentasikan kesalahan dalam bentuk persentase sehingga lebih mudah diinterpretasikan dan dibandingkan lintas model [8][22]. Semakin kecil nilai ketiga metrik tersebut, semakin baik kinerja model dalam meramalkan pergerakan indeks. Perbandingan nilai metrik antara model SARIMA, LSTM, dan GRU digunakan untuk menentukan model dengan akurasi peramalan tertinggi terhadap indeks harga saham S&P 500, sejalan dengan temuan pada beberapa penelitian komparasi model time series pada data keuangan sebelumnya [15][21][23][25][26][27].

3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk membandingkan performa metode SARIMA, GRU, dan LSTM untuk memprediksi harga penutupan (*close*) indeks S&P 500 (SPX). Tahapan *preprocessing* pada penelitian ini meliputi empat langkah utama, yaitu: (1) pemilihan fitur *close* sebagai target prediksi, (2) pembatasan periode observasi, (3) penyeragaman frekuensi data menjadi hari bursa (*business day*) dengan imputasi *forward-fill*, dan (4) pembagian data latih dan data uji dengan rasio 80:20 berbasis urutan waktu. Untuk model *deep learning* (GRU dan LSTM) yang sensitif terhadap skala fitur, dilakukan normalisasi *MinMaxScaler* dengan rentang [0, 1]. *Scaler difit* hanya pada data latih untuk mencegah kebocoran data (*data leakage*), kemudian diterapkan ke data latih dan data uji. Untuk model SARIMA, data asli (tanpa *scaling*) digunakan sesuai dengan asumsi model statistik berbasis *time series*. Sebelum melakukan pemodelan SARIMA, perlu dilakukan uji stasioneritas terhadap data latih karena SARIMA mensyaratkan data (atau hasil *differencing*-nya) bersifat stasioner. Uji stasioneritas yang digunakan adalah *Augmented Dickey-Fuller Test* dengan hipotesis nol (H_0) bahwa data mengandung unit *root* (tidak stasioner). Jika $p\text{-value} \leq 0,05$, maka H_0 ditolak dan data dinyatakan stasioner. Hasil Uji *Augmented Dickey-Fuller* (ADF) pada data latih disajikan pada Tabel 1.

Tabel 1. Hasil Uji *Augmented Dickey-Fuller* (ADF) pada Data Latih

Data	ADF Statistic	P-value	Kesimpulan
Close Asli	0,7991	0,9916	Tidak Stasioner (gagal tolak H_0)
Close Differencing-1	-42,6124	0,0000	Stasioner (tolak H_0)

Tabel 1 menunjukkan bahwa data *close* asli memiliki ADF statistic sebesar 0,7991 dengan $p\text{-value}$ 0,9916 ($> 0,05$), sehingga hipotesis nol gagal ditolak dan data dinyatakan tidak stasioner. Setelah dilakukan *differencing orde-1*, nilai ADF statistic turun drastis menjadi -42,6124 dengan $p\text{-value}$ 0,0000, sehingga data hasil *differencing* dinyatakan stasioner.

Temuan ini menjadi dasar pemilihan parameter *differencing* (d) pada model SARIMA, sekaligus menunjukkan adanya tren deterministik maupun stokastik pada data harga saham. Setelah dilakukan *preprocessing* data, uji stasioneritas, kemudian dilakukan pemodelan masing-masing algoritma, serta evaluasi perbandingan performa ketiga metode berdasarkan metrik MAE, RMSE, MAPE, dan R² Score sebagai berikut.

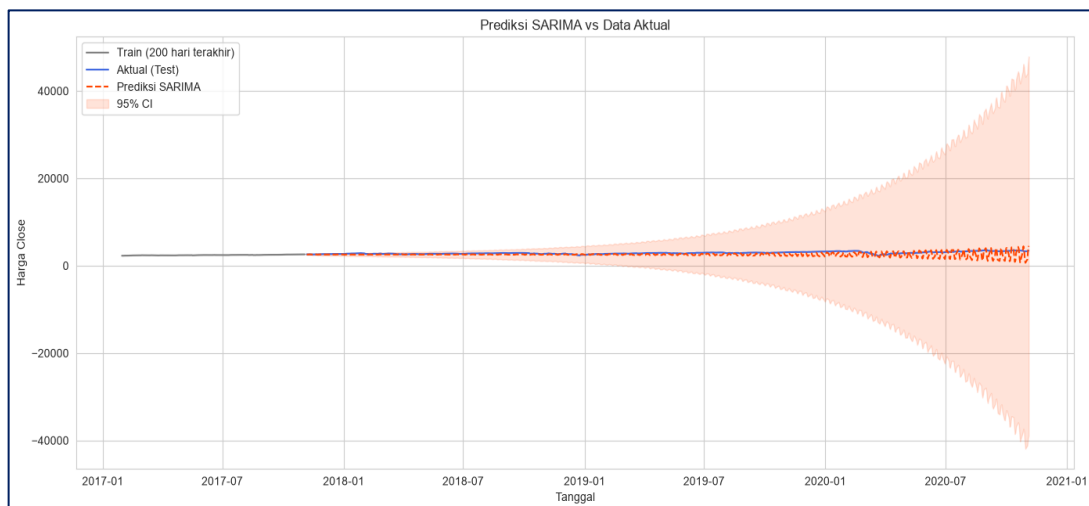
3.1 Pemodelan SARIMA

Pemodelan SARIMA dilakukan dengan bantuan *pmdarima.auto_arima* untuk menentukan orde (p, d, q) dan (P, D, Q)s secara otomatis berdasarkan nilai AIC minimum. Periode musiman (m) disetel ke 5 karena data merupakan data harian bursa dengan pola mingguan pada hari efektif (Senin-Jumat). Ruang pencarian yang digunakan adalah p, q ∈ {0, 1, 2, 3, 4, 5}, P, Q ∈ {0, 1, 2}, serta D = 1. Hasil penelusuran *stepwise* dari *auto_arima* pada tabel 2 menunjukkan model terbaik adalah SARIMA(4, 0, 2)x(2, 1, 0, 5) dengan nilai AIC sebesar 26.463,436 (waktu komputasi total sekitar 171,88 detik). Ringkasan parameter model SARIMA terpilih disajikan pada Tabel 2.

Tabel 2. Ringkasan Parameter Model SARIMA Terpilih

Parameter	Nilai
Order non-musiman (p, d, q)	(4, 0, 2)
Order musiman (P, D, Q, s)	(2, 1, 0, 5)
AIC	26.463,436
Log Likelihood	-13.221,718
BIC	26.523,911
HQIC	26.485,143

Selanjutnya model SARIMA dilatih ulang dengan parameter hasil *auto_arima* menggunakan fungsi SARIMA dari *statsmodels*. Ringkasan model menunjukkan bahwa seluruh koefisien signifikan pada taraf $\alpha = 0,05$ ($p\text{-value} < 0,05$) kecuali *intercept* ($p\text{-value}$ 0,119). Koefisien musiman ar.S.L5 (-0,6195) dan ar.S.L10 (-0,3721) keduanya signifikan, mengindikasikan adanya autokorelasi musiman pada lag 5 dan 10 (pola mingguan dan dua mingguan hari bursa). Hasil prediksi SARIMA pada periode data uji (783 hari bursa) kemudian dibandingkan dengan data aktual seperti yang ditunjukkan pada Gambar 2. Hasil tersebut menunjukkan bahwa model SARIMA gagal mengikuti perubahan harga aktual yang memiliki tren naik tajam pada periode 2019-2020, melainkan menghasilkan prediksi yang cenderung mendatar di sekitar rentang nilai harga awal periode.



Gambar 2. Prediksi SARIMA dan Data Aktual pada Data Uji

3.2 Pemodelan Gated Recurrent Unit (GRU)

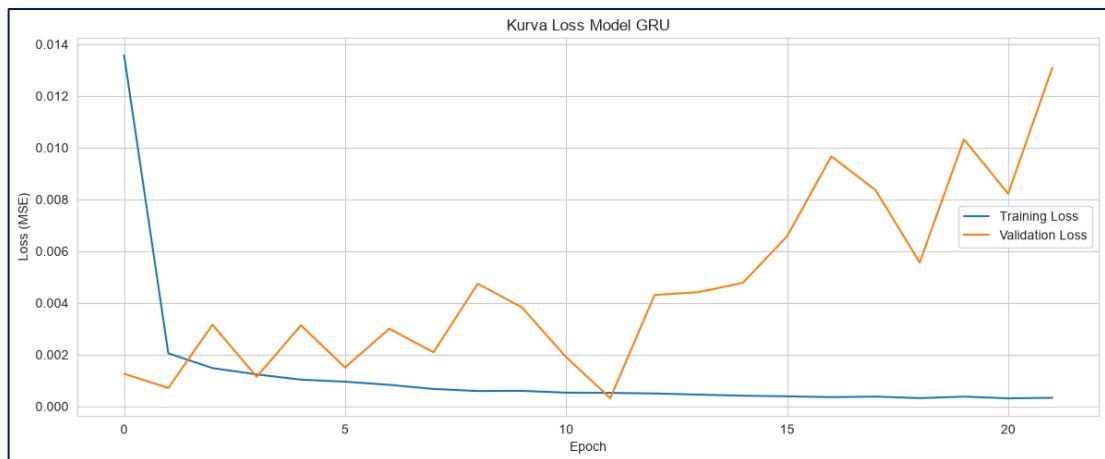
Model GRU dibangun dengan arsitektur dua layer GRU yang ditingkatkan dengan *dropout* untuk mencegah *overfitting*, serta dua layer *dense* di bagian akhir sebagai *regresor*. Lapisan pertama adalah GRU dengan 64 unit dan parameter *return_sequences=True*, yang berarti lapisan ini mengembalikan *output* pada setiap *time step* (bukan hanya *output* akhir). Hal ini terlihat dari *output shape* pada tabel 3 (None, 60, 64), di mana 60 merepresentasikan panjang jangka waktu (*time steps* atau *lookback window*) dan 64 adalah jumlah unit GRU. Lapisan ini memiliki 12.864 parameter, jumlah terbesar dalam model, karena harus memproses seluruh urutan data secara penuh. Setelah itu, lapisan *Dropout* dengan rate 0,2 diterapkan untuk mengurangi risiko *overfitting* dengan menonaktifkan 20% neuron secara acak selama pelatihan. Lapisan ini tidak memiliki parameter yang dilatih (0 parameter) karena hanya berfungsi sebagai mekanisme regularisasi, bukan transformasi bobot. Lapisan kedua GRU menggunakan 32 unit dengan *return_sequences=False*, sehingga hanya mengeluarkan *output* dari *time step* terakhir saja, ditunjukkan oleh perubahan *output shape* menjadi (None, 32). Lapisan ini "meringkas" seluruh informasi temporal yang telah diproses menjadi satu representasi vektor, dengan 9.408 parameter.

Setelah proses *sequencing*, dimensi data latih menjadi (3.071, 60, 1) dan data uji menjadi (783, 60, 1), di mana sumbu kedua adalah jumlah *timestep* dan sumbu ketiga adalah jumlah fitur. Arsitektur model GRU disajikan pada tabel 3.

Tabel 3. Arsitektur Model GRU

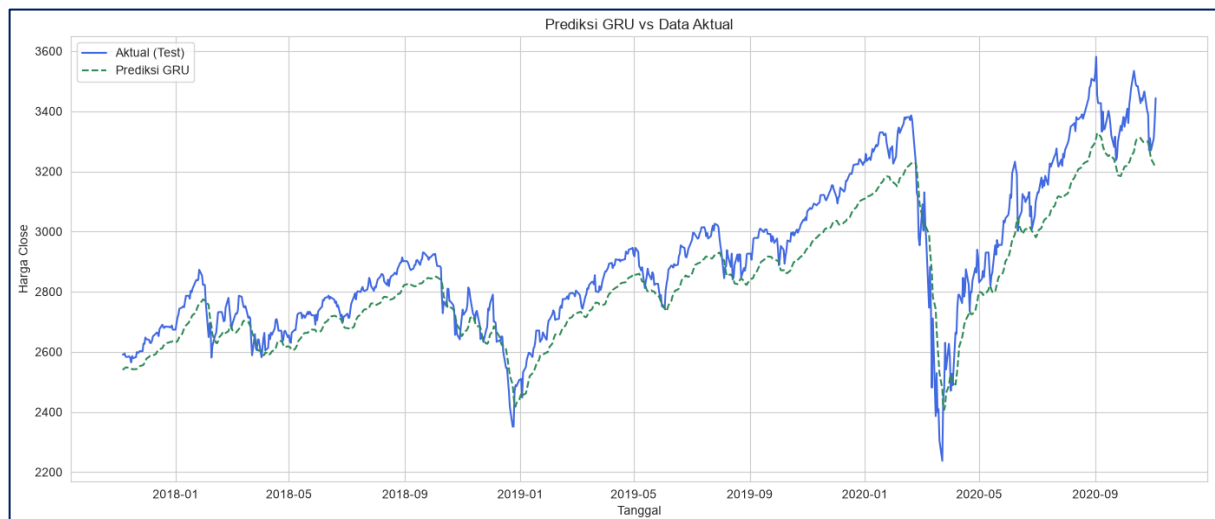
Layer (Tipe)	Output Shape	Jumlah Parameter
GRU (64 unit, <i>return_sequences=True</i>)	(None, 60, 64)	12.864
<i>Dropout</i> (0,2)	(None, 60, 64)	0
GRU (32 unit, <i>return_sequences=False</i>)	(None, 32)	9.408
<i>Dropout</i> (0,2)	(None, 32)	0
<i>Dense</i> (16 unit, ReLU)	(None, 16)	528
<i>Dense</i> (1 unit, linier)	(None, 1)	17
Total	-	22.817

Kurva *Loss* Pelatihan dan Validasi Model GRU disajikan pada Gambar 3.



Gambar 3. Kurva *Loss* Pelatihan dan Validasi Model GRU

Perbandingan hasil prediksi GRU dan data aktual pada data uji disajikan pada Gambar 4.



Gambar 4. Prediksi GRU dan Data Aktual pada Data Uji

Model dikompilasi dengan *optimizer* Adam dan *loss function* Mean Squared Error (MSE), serta metrik tambahan Mean Absolute Error (MAE). Pelatihan dilakukan selama maksimum 100 *epoch* dengan *batch size* 16 dan *validation split* 10% dari data latih. Mekanisme *EarlyStopping* dengan *patience* 10 dan *restore_best_weights=True* diterapkan untuk menghentikan pelatihan saat *val_loss* tidak kunjung membaik dan mengembalikan bobot model terbaik. Pelatihan berhenti pada *epoch* ke-22 dengan *val_loss* minimum. Kurva *loss* pelatihan dan validasi ditunjukkan pada Gambar 3. Terlihat bahwa *loss* pelatihan menurun secara konsisten dari 0,0136 (*epoch* 1) menjadi sekitar $3,3 \times 10^{-4}$ (*epoch* 22), sementara *loss* validasi menurun lebih fluktuatif dan menyentuh nilai minimum $3,19 \times 10^{-4}$ pada *epoch* ke-12. Setelah titik tersebut, *val_loss* cenderung naik sehingga *EarlyStopping* mengembalikan bobot *epoch* terbaik. Setelah pelatihan, model GRU digunakan untuk memprediksi data uji. Hasil prediksi (yang sebelumnya berada pada skala [0, 1]) kemudian di *inverse transform* kembali ke skala harga asli menggunakan *scaler* yang sama. Visualisasi prediksi GRU versus data

aktual ditunjukkan pada Gambar 4, dimana terlihat bahwa pola data hasil prediksi model GRU memiliki pola yang hampir dengan pola data aktua.

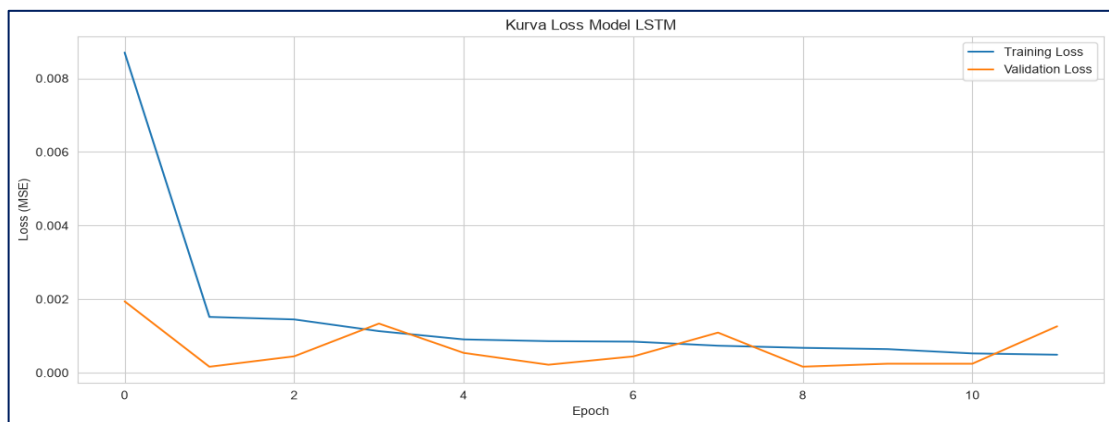
3.3 Pemodelan Long Short-Term Memory (LSTM)

Pemodelan LSTM dilakukan dengan struktur sekuens yang serupa dengan GRU (WINDOW_SIZE = 60), namun menggunakan dua layer LSTM yang masing-masing memiliki mekanisme *memory cell* dan gerbang (*input, forget, output*) yang dirancang untuk menangani dependensi jangka panjang. Lapisan pertama LSTM dapat dilihat pada tabel 4 dengan 64 unit dan parameter *return_sequences=True*, menghasilkan output berdimensi (None, 60, 64) dengan 16.896 parameter. Penggunaan *return_sequences=True* pada lapisan ini penting karena *output* sekuensial (bukan hanya *output* akhir) diperlukan sebagai input bagi lapisan LSTM berikutnya, sehingga informasi temporal dari seluruh 60 *time step* (misalnya representasi 60 hari data historis) tetap dipertahankan. Lapisan *Dropout* (0,2) diterapkan setelahnya untuk mengurangi risiko *overfitting* dengan menonaktifkan secara acak 20% *neuron* selama proses pelatihan, tanpa menambah parameter baru. Lapisan kedua adalah LSTM dengan 32 unit dan *return_sequences=False*, yang menghasilkan *output* berdimensi (None, 32) dengan 12.416 parameter. Karena *return_sequences=False*, lapisan ini hanya mengeluarkan *output* dari *time step* terakhir, meringkas seluruh informasi sekuensial menjadi satu vektor representasi berdimensi 32 yang siap diteruskan ke lapisan *fully-connected*. Arsitektur model LSTM ditunjukkan pada Tabel 4.

Tabel 4. Arsitektur Model LSTM

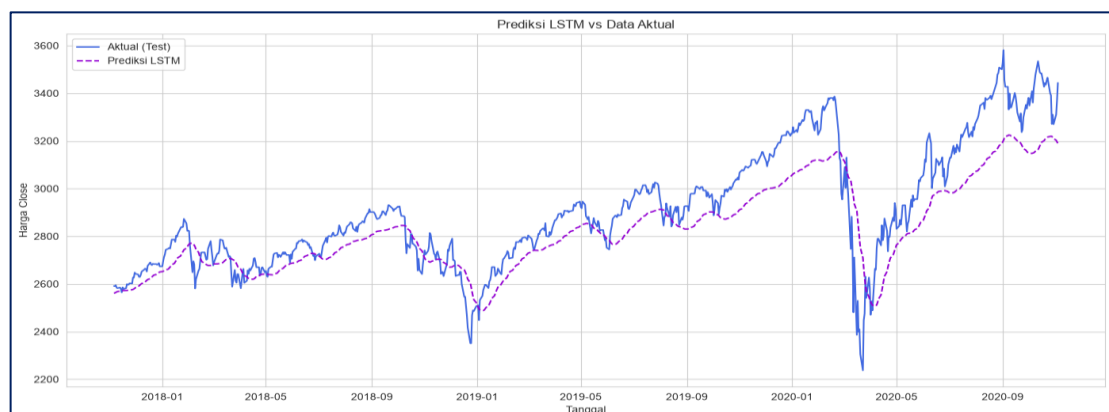
Layer (Tipe)	Output Shape	Jumlah Parameter
LSTM (64 unit, <i>return_sequences=True</i>)	(None, 60, 64)	16.896
Dropout (0,2)	(None, 60, 64)	0
LSTM (32 unit, <i>return_sequences=False</i>)	(None, 32)	12.416
Dropout (0,1)	(None, 32)	0
Dense (16 unit, ReLU)	(None, 16)	528
Dense (1 unit, linier)	(None, 1)	17
Total	-	29.857

Kurva *Loss* Pelatihan dan Validasi Model LSTM disajikan pada Gambar 5.



Gambar 5. Kurva *Loss* Pelatihan dan Validasi Model LSTM

Perbandingan hasil prediksi LSTM dan data aktual pada data uji disajikan pada Gambar 6.



Gambar 6. Prediksi LSTM dan Data Aktual pada Data Uji

Pelatihan menggunakan konfigurasi yang sama dengan GRU (*optimizer* Adam, *loss* MSE, *batch size* 16, maksimum 100 *epoch*, *validation split* 10%, *EarlyStopping patience* 10). Terlihat pada Gambar 5 pelatihan berhenti pada *epoch* ke-12 dengan *val_loss minimum* sebesar $1,64 \times 10^{-4}$ pada *epoch* ke-9. Dibandingkan GRU, LSTM berhenti lebih awal, namun memiliki jumlah parameter yang lebih besar (29.857 dibanding 22.817) karena setiap unit LSTM memiliki empat gerbang (*forget*, *input*, *cell*, *output*) sedangkan GRU hanya memiliki tiga gerbang (*reset*, *update*, *candidate*). Prediksi LSTM pada data uji di *inverse transform* kembali ke skala harga asli, kemudian dibandingkan dengan data aktual seperti yang ditunjukkan pada Gambar 6. Secara visual, prediksi LSTM mampu mengikuti pola umum data aktual, namun tertinggal (*lag*) pada titik-titik perubahan harga yang tajam.

3.4 Evaluasi dan Perbandingan Performa Model

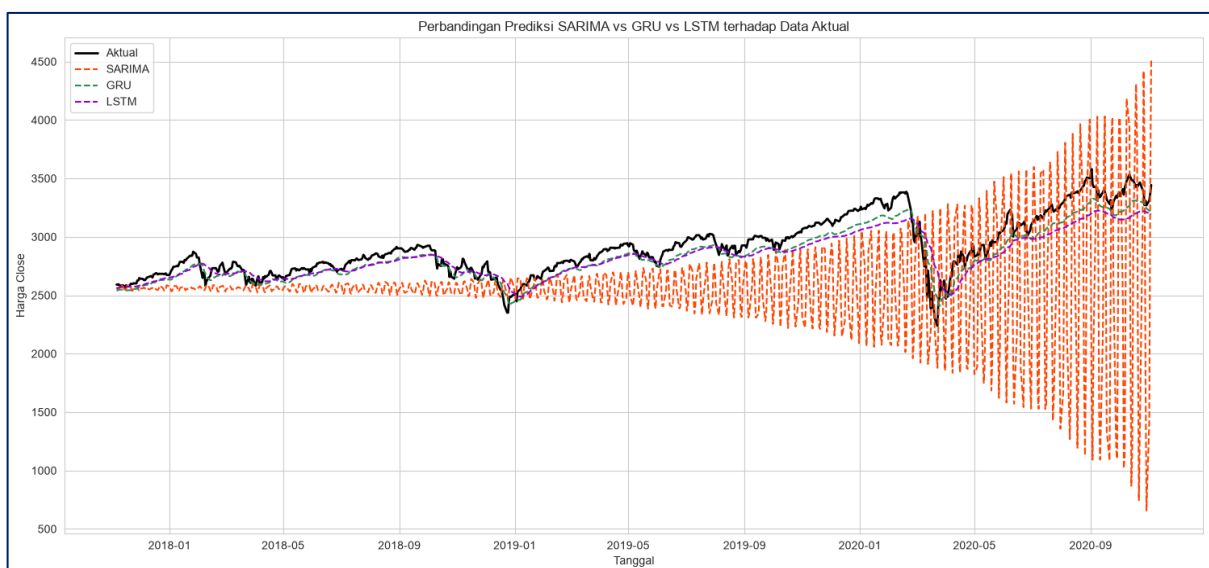
Evaluasi kinerja ketiga model dilakukan pada rentang tanggal yang sama, yaitu mengikuti panjang prediksi GRU dan LSTM (783 hari bursa). Hal ini dilakukan karena *sekuens input* memerlukan *WINDOW_SIZE* = 60 observasi sebelumnya, sehingga 60 observasi pertama pada data uji tidak memiliki prediksi. Metrik evaluasi yang digunakan adalah *Mean Absolute Error* (MAE), *Root Mean Squared Error* (RMSE), *Mean Absolute Percentage Error* (MAPE), dan *R² Score*. Hasil evaluasi ketiga model ditunjukkan pada Tabel 5.

Tabel 5. Perbandingan Metrik Evaluasi Model SARIMA, GRU, dan LSTM

Model	MAE	RMSE	MAPE (%)	R ² Score
SARIMA	418,7748	610,1970	13,7166	-5,0196
GRU	87,0388	100,7213	2,9387	0,8360
LSTM	105,8862	132,1261	3,5476	0,7178

Berdasarkan Tabel 5 model GRU menghasilkan performa terbaik dengan nilai MAE terendah (87,04), RMSE terendah (100,72), MAPE terendah (2,94%), serta R² Score tertinggi (0,836). Model LSTM menempati posisi kedua dengan MAE 105,89, RMSE 132,13, MAPE 3,55%, dan R² Score 0,718. Sementara itu, model SARIMA menunjukkan performa yang jauh lebih buruk dengan MAE 418,77, RMSE 610,20, MAPE 13,72%, dan R² Score negatif (-5,02). Nilai R² yang negatif mengindikasikan bahwa model SARIMA berperforma kurang baik dibandingkan prediksi naif berbasis rata-rata, hal ini mengisyaratkan bahwa asumsi linearitas dan stasioneritas pada model SARIMA tidak cukup untuk menangkap pola non-linier dan tren kuat pada data S&P 500.

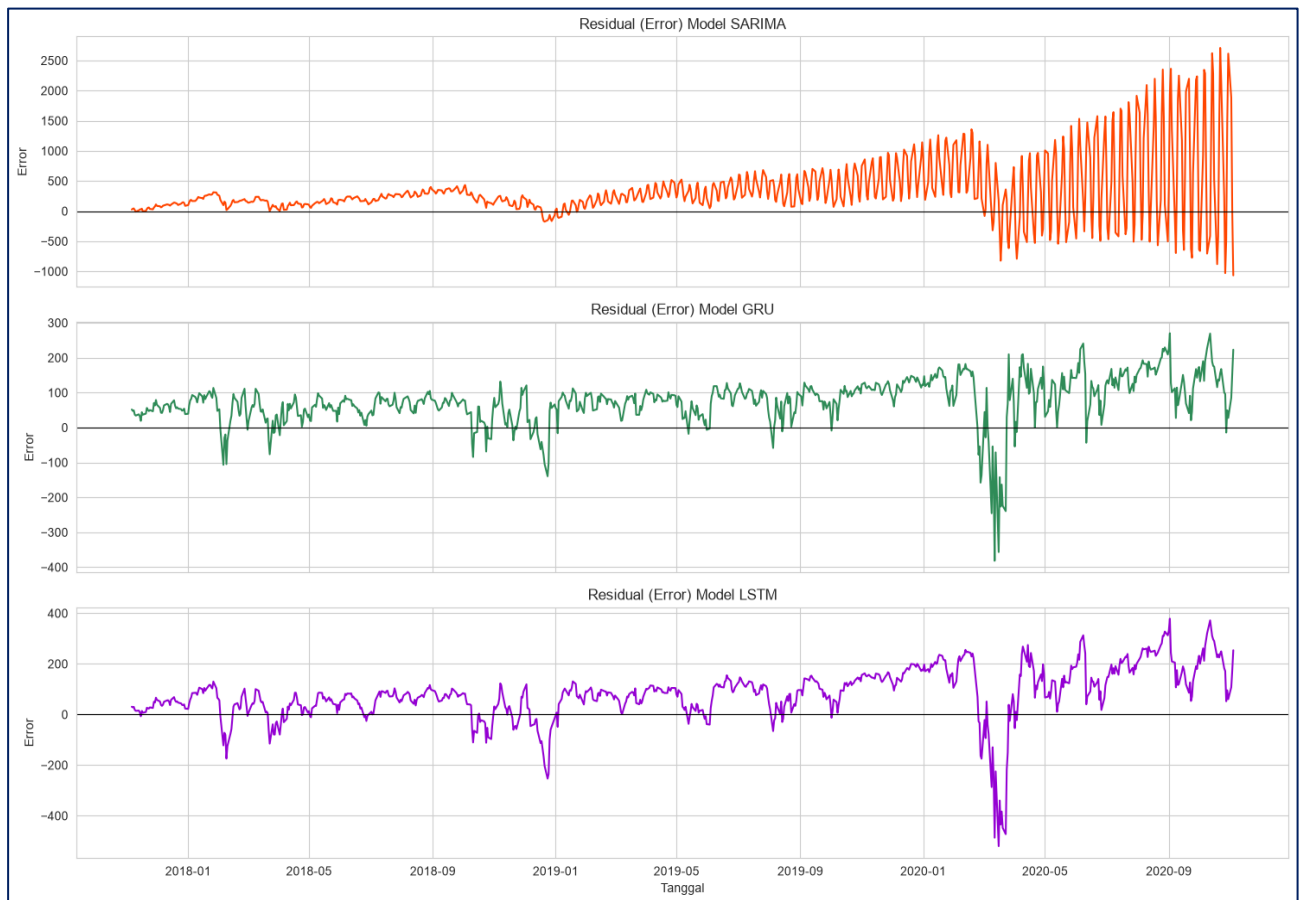
Berdasarkan pada Gambar 7 terlihat bahwa kurva prediksi GRU dan LSTM hampir berhimpitan dengan kurva aktual, sementara kurva prediksi SARIMA tampak jauh di bawah nilai aktual pada paruh kedua periode prediksi. Hal ini konsisten dengan hasil numerik pada Tabel 5, di mana kedua model *deep learning* menghasilkan *error* yang jauh lebih kecil dibanding SARIMA. Gambar 7 menampilkan grafik residual (*error*) masing-masing model. Residual SARIMA menunjukkan pola sistematis yang melebar ke arah negatif (cenderung *under-estimate*) seiring naiknya nilai aktual, hal ini merupakan indikasi klasik bahwa model linear tidak mampu menangkap tren non-linear. Sebaliknya, residual GRU dan LSTM tampak lebih stasioner di sekitar nol dengan amplitudo yang jauh lebih kecil. *Overlay* prediksi ketiga model terhadap data aktual ditampilkan pada Gambar 7.



Gambar 7. Overlay Prediksi SARIMA, GRU, dan LSTM terhadap Data Aktual

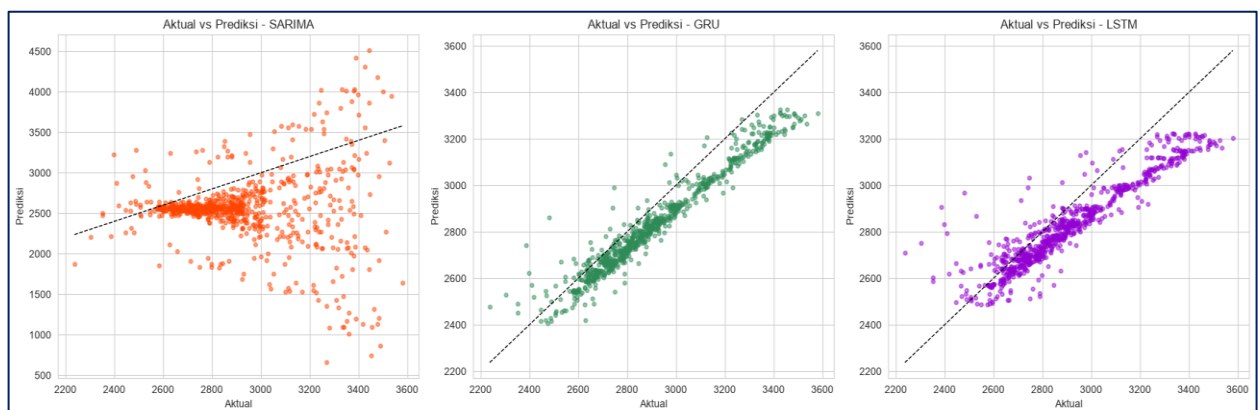
Secara umum perbandingan residual ketiga model pada Gambar 8 menegaskan bahwa model berbasis *deep learning* (GRU dan LSTM) memberikan performansi lebih baik dibandingkan SARIMA dalam menangani data deret waktu harga saham yang bersifat non-linear dan volatil. SARIMA menunjukkan residual yang membesar secara sistematis seiring waktu yang mengindikasikan model misspecification (ketidakmampuan menangkap heteroskedastisitas),

sedangkan GRU dan LSTM mempertahankan residual yang jauh lebih terkontrol di sekitar nol, meskipun keduanya tetap menunjukkan sensitivitas terhadap peristiwa shock ekstrem seperti awal pandemi COVID-19. Di antara GRU dan LSTM, GRU memiliki sebaran residual yang sedikit lebih sempit dan stabil, yang konsisten dengan temuan bahwa GRU memberikan akurasi terbaik dalam penelitian ini. Perbandingan tren residual ketiga model disajikan pada Gambar 8.



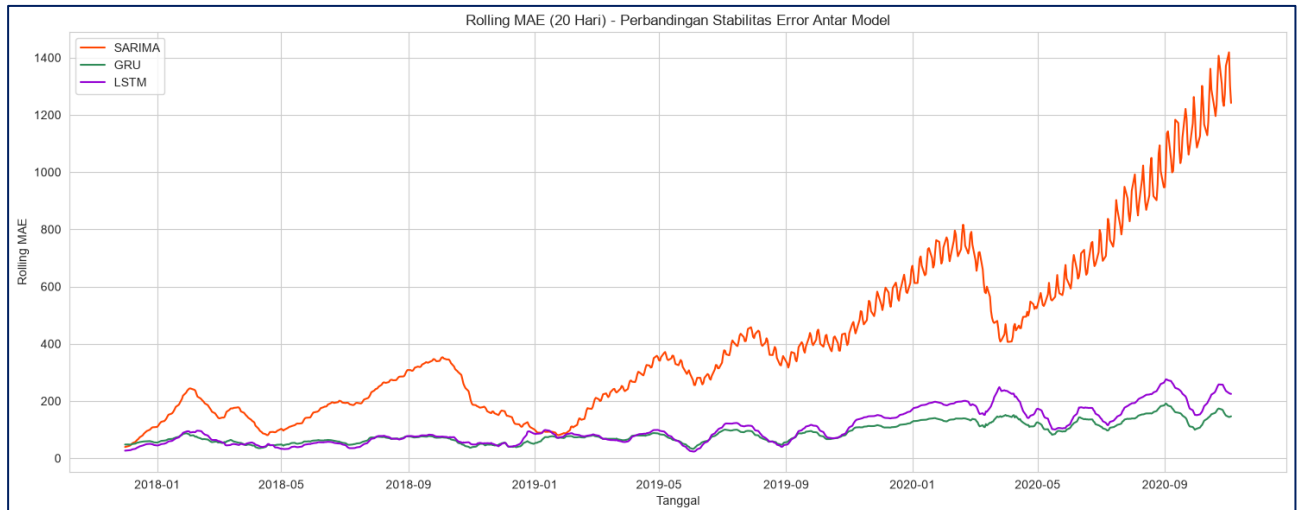
Gambar 8. Grafik Residual (*Error*) Masing-Masing Model

Scatter plot data aktual dan prediksi pada Gambar 9 memperlihatkan seberapa dekat prediksi masing-masing model terhadap garis diagonal $y = x$. Titik-titik prediksi SARIMA berada di bawah garis diagonal karena nilai aktual terus meningkat sementara prediksi SARIMA cenderung datar. Titik-titik prediksi GRU paling rapat di sekitar garis diagonal, mengindikasikan akurasi paling tinggi. LSTM juga menunjukkan pola yang positif namun dengan sebaran yang sedikit lebih lebar dibanding GRU. Gambar 9 menampilkan *rolling* MAE (jendela 20 hari) yang merepresentasikan stabilitas *error* dari waktu ke waktu. Kurva *rolling* MAE GRU konsisten berada di bawah kedua model lainnya pada hampir seluruh horizon prediksi, mengindikasikan stabilitas prediksi yang lebih baik. *Rolling* MAE LSTM berada sedikit di atas GRU, namun tetap berada di bawah SARIMA dengan selisih yang sangat besar. *Scatter plot* data aktual dan prediksi ketiga model disajikan pada Gambar 9.



Gambar 9. *Scatter Plot* Data Aktual dan Prediksi Masing-Masing Model

Rolling MAE yang digunakan untuk mengukur stabilitas residual ketiga model disajikan pada Gambar 10.



Gambar 10. Rolling MAE (20 Hari) untuk Ketiga Model

Salah satu ukuran kebaikan model peramalan dapat dilihat dari visualisasi nilai residual ketiga model yang stabil atau tidak membentuk tren tertentu (*Rolling MAE*) pada gambar 10. Performa SARIMA yang kurang baik pada penelitian ini dapat dijelaskan oleh beberapa faktor. Pertama, data harga saham S&P 500 pada 15 tahun terakhir memiliki tren naik yang kuat dan mengandung perubahan struktural (misalnya saat pandemi COVID-19 pada awal 2020) yang sulit ditangkap oleh model linear. Kedua, meski *auto_arima* memilih model dengan AIC minimum pada data latih, pemilihan orde tersebut belum tentu optimal untuk horizon prediksi panjang. Ketiga, SARIMA bekerja pada data asli (tanpa *differencing* internal untuk prediksi *out-of-sample*) sehingga prediksi bertumbuh secara eksponensial berdasarkan orde integrasi, namun dalam praktiknya gagal menangkap perubahan tingkat harga.

Sementara itu, GRU mengungguli LSTM pada eksperimen ini dengan selisih yang cukup signifikan. Hal ini menarik karena dalam banyak literatur LSTM dianggap lebih baik untuk dependensi jangka panjang [17][18]. Beberapa faktor yang dapat menjelaskan fenomena ini adalah: (1) ukuran data yang relatif terbatas (3.071 sekuens latih) membuat LSTM dengan parameter lebih banyak (29.857) lebih rentan terhadap *overfitting* meskipun telah diterapkan *Dropout*; (2) GRU memiliki arsitektur yang lebih sederhana sehingga lebih cepat konvergen pada dataset berukuran menengah; (3) *WINDOW SIZE* = 60 hari (± 3 bulan) tidak terlalu panjang, sehingga keunggulan LSTM dalam menangkap dependensi jangka panjang tidak terlalu kentara. Terlepas dari keunggulan GRU pada eksperimen ini, baik GRU maupun LSTM masih menunjukkan kelemahan berupa *lagging prediction*, yaitu prediksi yang cenderung mengikuti harga satu langkah sebelumnya ($t = -1$) sehingga kurang mampu mengantisipasi perubahan harga yang tajam. Hal ini merupakan keterbatasan umum model *sequence-to-one* yang dapat diatasi dengan pendekatan *multi-step*, penambahan fitur eksogen (misalnya volume, indikator teknikal, atau data makroekonomi), maupun penggunaan arsitektur *encoder-decoder* dengan mekanisme *attention*. Secara umum hasil dari penelitian ini memperkuat temuan penelitian sebelumnya [17][18] bahwa model *deep learning* khususnya LSTM dan GRU lebih baik dibandingkan model statistika klasik yang dalam penelitian ini menggunakan SARIMA dalam meramalkan data *non-linear* pada data harga penutupan harian indeks S&P 500 dari dataset publik. Penelitian ini menunjukkan bahwa model GRU sesuai digunakan untuk data *non-linear* dengan jumlah data relatif terbatas karena memberikan perormasi terbaik dibandingkan LSTM dan SARIMA.

3.5 Pembahasan

Jika dibandingkan dengan penelitian terdahulu, temuan penelitian ini menunjukkan pola yang sebagian sejalan namun juga memperlihatkan beberapa perbedaan penting. Keunggulan model *deep learning* dibandingkan model statistik klasik pada penelitian ini konsisten dengan temuan Kirelli [14] yang melaporkan LSTM lebih unggul dibandingkan ARIMA pada harga saham perusahaan teknologi, maupun Hossain et al. [16] yang menunjukkan LSTM mengungguli ARIMA secara signifikan pada sektor keuangan S&P 500. Penurunan MAPE dari 13,72% (SARIMA) menjadi 2,94% (GRU) pada penelitian ini bahkan melampaui besaran perbaikan yang dilaporkan Hossain et al. [16], yang membatasi perbaikan RMSE dan MAE masing-masing pada kisaran 15% dan 30%. Kesenjangan performa yang lebih lebar pada penelitian ini kemungkinan besar disebabkan oleh rentang data historis yang lebih panjang (15 tahun) sehingga mencakup instabilitas struktural seperti pandemi COVID-19, kondisi yang secara khusus menyulitkan asumsi linearitas SARIMA.

Terkait perbandingan antara LSTM dan GRU, temuan bahwa GRU sedikit lebih unggul dibandingkan LSTM pada penelitian ini sejalan dengan hasil Berradi et al. [17] yang juga membandingkan LSTM dan GRU pada data harian S&P 500, meskipun pada rentang waktu observasi yang berbeda (1991-2021). Namun demikian, temuan ini berbeda dengan Li dan Chen [18] yang melaporkan LSTM tunggal memberikan performa terbaik dibandingkan GRU maupun hybrid LSTM-GRU pada objek kajian indeks S&P 500 yang sama. Perbedaan ini kemungkinan besar bersumber dari ukuran data latih dan konfigurasi arsitektur yang digunakan. Penelitian ini menggunakan data latih berjumlah 3.071 sekuens yang relatif terbatas, sehingga LSTM dengan jumlah parameter lebih besar (29.857) lebih rentan mengalami *overfitting* dibandingkan GRU yang lebih ringkas (22.817 parameter). Temuan ini memperkuat kesimpulan Şengül Abedi [15] bahwa



tidak ada satu model yang selalu unggul secara universal, karena performa GRU maupun LSTM sangat bergantung pada karakteristik data, ukuran sampel, dan konfigurasi eksperimen.

Dari sisi kebaruan, penelitian ini melengkapi kesenjangan yang teridentifikasi pada tinjauan pustaka, yaitu terbatasnya studi yang membandingkan SARIMA, LSTM, dan GRU secara simultan pada data historis indeks S&P 500. Studi tiga model oleh Şengül Abedi [15] dilakukan pada domain perdagangan luar negeri sektor pertanian dan pangan, sehingga karakteristik datanya berbeda dari data indeks harga saham yang memiliki tren jangka panjang dan volatilitas yang lebih tinggi. Penelitian ini juga memberikan wawasan tambahan mengenai *trade off* antara akurasi dan kompleksitas parameter LSTM-GRU yang menurut Akbari et al. [13] masih relatif jarang dikaji secara eksplisit pada konteks prediksi indeks S&P 500, dengan menunjukkan bahwa arsitektur yang lebih sederhana (GRU) tidak selalu mengorbankan akurasi, bahkan dapat memberikan hasil yang lebih baik pada kondisi data yang relatif terbatas. Dengan demikian, hasil penelitian ini secara umum memperkuat dan memperluas temuan penelitian-penelitian terdahulu, sekaligus menegaskan pentingnya mempertimbangkan karakteristik data dan konteks kajian dalam memilih model peramalan deret waktu keuangan yang paling sesuai.

4. KESIMPULAN

Penelitian ini berhasil menjawab tujuan yang ditetapkan, yaitu merancang, melatih, dan membandingkan tiga model peramalan harga penutupan harian indeks S&P 500 secara kuantitatif menggunakan data historis dari Kaggle. Proses pemodelan mengonfirmasi bahwa data indeks S&P 500 tidak stasioner pada level aslinya dan baru menjadi stasioner setelah proses *differencing orde* pertama, sehingga model SARIMA(4,0,2)(2,1,0,5) hasil pencarian otomatis *auto_arima* dipilih sebagai representasi pendekatan statistik klasik. Perbandingan kuantitatif pada 783 hari bursa data uji membuktikan bahwa kedua model *deep learning*, GRU dan LSTM, secara konsisten unggul jauh di atas SARIMA pada seluruh metrik evaluasi, dengan penurunan MAPE dari 13,72% pada SARIMA menjadi 2,94% pada GRU dan 3,55% pada LSTM, sekaligus mengubah R^2 Score dari negatif (-5,02) pada SARIMA menjadi positif dan tinggi pada GRU (0,836). Temuan ini secara empiris menegaskan bahwa asumsi linearitas dan stasioneritas pada model statistik klasik tidak memadai untuk merepresentasikan dinamika harga saham yang *non-linear*, *volatil*, dan mengandung perubahan struktural seperti guncangan pandemi COVID-19. Di antara dua model *deep learning*, GRU terbukti lebih unggul dibandingkan LSTM pada konteks data dan ukuran sampel penelitian ini (3.071 sekuens data latih), dengan arsitektur yang lebih ringkas (22.817 parameter berbanding 29.857 parameter) namun menghasilkan akurasi dan stabilitas *rolling* MAE yang lebih baik. Kondisi ini mengindikasikan bahwa pada skenario dengan volume data terbatas dan window historis menengah (60 hari), kesederhanaan struktur gerbang GRU memberikan keseimbangan yang lebih baik antara kapasitas model dan risiko *overfitting* dibandingkan kompleksitas tambahan yang ditawarkan LSTM. Meski demikian, kedua model *deep learning* tetap menunjukkan kelemahan berupa *lagging prediction* terhadap pergerakan harga yang tajam, sehingga akurasi tinggi yang dicapai belum sepenuhnya menjamin ketepatan pada titik pembalikan tren. Secara keseluruhan, penelitian ini memberikan bukti empiris baru mengenai keunggulan komparatif GRU sebagai model peramalan yang efisien secara komputasi sekaligus akurat untuk indeks pasar modal global berskala besar seperti S&P 500, sekaligus melengkapi kesenjangan penelitian terdahulu yang jarang membandingkan SARIMA, LSTM, dan GRU secara simultan pada objek kajian yang sama. Hasil ini menjadi dasar rekomendasi praktis bagi investor dan analis untuk mempertimbangkan GRU sebagai alternatif model peramalan yang lebih hemat sumber daya tanpa mengorbankan akurasi. Penelitian ini memiliki beberapa keterbatasan yang dapat dieksplorasi pada penelitian selanjutnya, antara lain: (1) pencarian *hyperparameter* model dilakukan secara manual sehingga belum tentu optimal; (2) fitur *input* hanya menggunakan harga *Close* tanpa menyertakan Volume atau variabel eksogen lain; (3) evaluasi dilakukan hanya pada satu skema split train-test tanpa validasi berbasis walk-forward cross-validation; (4) arsitektur yang diuji masih relatif sederhana tanpa membandingkan varian seperti *Bidirectional* GRU atau LSTM atau *stacked encoder-decoder*; serta (5) horizon prediksi hanya satu hari ke depan sehingga belum menggambarkan kinerja model untuk prediksi *multi-step*. Penelitian lanjutan disarankan melakukan *tuning hyperparameter* secara sistematis, menambahkan fitur eksogen seperti Volume atau VIX, menerapkan skema validasi *walk-forward*, serta mengeksplorasi arsitektur *hybrid* seperti CNN-LSTM atau *Transformer* untuk memperkuat ketahanan model terhadap guncangan pasar yang ekstrem.

REFERENCES

- [1] S&P Dow Jones Indices, "S&P 500," S&P Global. [Online]. Available: <https://www.spglobal.com/spdji/en/indices/equity/sp-500/>. [Accessed: Aug. 05, 2026].
- [2] Yahoo Finance, "S&P 500 (^GSPC) Historical Data." [Online]. Available: <https://finance.yahoo.com/quote/%5EGSPC/history>. [Accessed: Aug. 05, 2026].
- [3] H. Yilmazkuday, "Geopolitical risk and stock prices," European Journal of Political Economy, vol. 83, p. 102553, 2024, doi: 10.1016/j.ejpoleco.2024.102553.
- [4] A. Shleifer, Inefficient Markets: An Introduction to Behavioral Finance. Oxford, U.K.: Oxford University Press, 2000.
- [5] Z. Wang, "Stock price prediction using LSTM neural networks: Techniques and applications," Applied and Computational Engineering, vol. 86, pp. 275-281, 2024, doi: 10.54254/2755-2721/86/20241605.
- [6] C. Nugent, "S&P 500 stock data," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/camnugent/sandp500>. [Accessed: Aug. 21, 2026].



- [7] O. S. Adesina and L. O. Obokoh, "A hybrid framework of deep learning and traditional time series models for exchange rate prediction," *Scientific African*, vol. 29, p. e02818, 2025, doi: 10.1016/j.sciaf.2025.e02818.
- [8] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021.
- [9] A. Nielsen, *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. Sebastopol, CA, USA: O'Reilly Media, 2019.
- [10] W. Chen, W. Hussain, F. Cauteruccio, and X. Zhang, "Deep learning for financial time series prediction: A state-of-the-art review of standalone and hybrid models," *Computer Modeling in Engineering & Sciences*, vol. 139, no. 1, pp. 187-224, 2024, doi: 10.32604/cmescs.2023.031388.
- [11] A. Farhadi, A. Zamanifar, A. Alipour, A. Taheri, and M. Asadolahi, "A hybrid LSTM-GRU model for stock price prediction," *IEEE Access*, vol. 13, pp. 117594-117618, 2025, doi: 10.1109/ACCESS.2025.3586558.
- [12] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
- [13] R. W. Akbari, A. P. Kuncoro, and A. Jahir, "Comparative study of LSTM and GRU accuracy in predicting BBRI stock closing price," *Journal of Applied Informatics and Computing*, vol. 10, no. 1, pp. 837-846, 2026, doi: 10.30871/jaic.v10i1.11938.
- [14] Y. Kirelli, "Comparative analysis of LSTM and ARIMA models in stock price prediction: A technology company example," *Black Sea Journal of Engineering and Science*, vol. 7, no. 5, pp. 866-873, Sep. 2024, doi: 10.34248/bsengineering.1445997.
- [15] Z. Şengül Abedi, "Forecasting foreign trade in Türkiye's agri-food sector: A comparative analysis of SARIMA, LSTM, and GRU models," *Turkish Journal of Agricultural and Natural Sciences*, vol. 12, no. 4, pp. 1007-1026, 2025, doi: 10.30910/turkjans.1756810.
- [16] M. A. Hossain, R. Karim, M. R. Thulasiram, and B. M. Nguyen, "Systematic comparison, evaluation and identification of robust model to forecast the closing price of S&P 500 financial sector through classical and AI-based approaches," *Journal of Risk and Financial Management*, vol. 16, no. 3, p. 165, 2023, doi: 10.3390/jrfm16030165.
- [17] S. Berradi, M. Lazaar, and H. H. H. Aldayel, "Forecasting the S&P 500 financial sector using deep learning: LSTM vs. GRU," *International Journal of Financial Innovation in Banking*, vol. 4, no. 1, pp. 1-18, 2023.
- [18] Y. Li and H. Chen, "A comparative study of LSTM, GRU, and hybrid LSTM-GRU models for forecasting S&P 500 index," *Highlights in Business, Economics and Management*, vol. 21, pp. 132-145, 2024, doi: 10.54097/hbem.v21i.13601.
- [19] A. Yunita, M. I. Pratama, M. Z. Almuzakki, H. Ramadhan, E. A. P. Akhir, A. B. F. Mansur, and A. H. Basori, "Performance analysis of neural network architectures for time series forecasting: A comparative study of RNN, LSTM, GRU, and hybrid models," *MethodsX*, vol. 15, p. 103462, 2025, doi: 10.1016/j.mex.2025.103462.
- [20] O. Moseane, J. T. Tsoku, and D. Metsileng, "Hybrid time series and ANN-based ELM model on JSE/FTSE closing stock prices," *Frontiers in Applied Mathematics and Statistics*, vol. 10, p. 1454595, 2024, doi: 10.3389/fams.2024.1454595.
- [21] K. Lahboub and M. Benali, "Assessing the predictive power of transformers, ARIMA, and LSTM in forecasting stock prices of Moroccan credit companies," *Journal of Risk and Financial Management*, vol. 17, no. 7, p. 293, 2024, doi: 10.3390/jrfm17070293.
- [22] P. Pilla and R. Mekonen, "Forecasting S&P 500 using LSTM models," *arXiv preprint arXiv:2501.17366*, 2025.
- [23] F. Kamalov, L. Smail, and I. Gurrib, "Forecasting with deep learning: S&P 500 index," in *Proc. 2020 13th Int. Symp. Computational Intelligence and Design (ISCID)*, 2020, pp. 422-425.
- [24] F. H. Sezgin, Ö. Algorabi, G. Sart, and M. Güler, "Hyperparameter-optimized RNN, LSTM, and GRU models for airline stock price prediction: A comparative study on THYAO and PGSUS," *Symmetry*, vol. 17, no. 11, p. 1905, 2025, doi: 10.3390/sym17111905.
- [25] M.-J. Park and H.-S. Yang, "Comparative study of time series analysis algorithms suitable for short-term forecasting in implementing demand response based on AMI," *Sensors*, vol. 24, no. 22, p. 7205, 2024, doi: 10.3390/s24227205.
- [26] M. L. Hossain, S. M. N. Shams, and S. M. Ullah, "Time-series and deep learning approaches for renewable energy forecasting in Dhaka: A comparative study of ARIMA, SARIMA, and LSTM models," *Discover Sustainability*, vol. 6, art. no. 775, 2025, doi: 10.1007/s43621-025-01733-5.
- [27] William, A. K. Wilham, and S. R. Manalu, "Stocks and models: A comparative benchmark of machine learning and deep learning approaches for stock market prediction," *Procedia Computer Science*, vol. 269, pp. 1532-1545, 2025, doi: 10.1016/j.procs.2025.09.095.