Stock Market Forecasting Using Sentiment Analysis: A Machine Learning and Deep Learning Hybrid Approach

Abstract

This study develops a hybrid forecasting system that integrates sentiment analysis, deep learning (DL) and machine learning (ML) to improve the forecasting of stock market indices. We focus on three key indices: the FTSE 100 (United Kingdom), the DAX (Germany) and the S&P 500 (United States). Traditional quantitative market indicators are combined with sentiment indicators derived from financial news headlines collected from Yahoo Finance between January 2020 and June 2023. Sentiment is measured with the FinBERT model and aggregated daily to match the market data. Four hybrid models (LSTM + RF, BiLSTM + RF, CNN + RF and DNN + RF) are developed to combine the temporal sequence learning of deep learning with the non-linear modelling capacity of Random Forest, and their performance is compared with that of the stand-alone models (RF, LSTM, BiLSTM, CNN and DNN). Evaluation with the root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2) and directional accuracy (DirAcc) shows that the hybrid models perform best: LSTM + RF attains the lowest RMSE of 1.97 and a DirAcc of 0.94 for the DAX with sentiment, compared with an RMSE of 14.59 for the stand-alone LSTM. The hybrid models consistently show low errors, high explanatory power and strong directional accuracy. Sentiment improves performance, particularly for the DAX, but has a limited and model-dependent effect for the FTSE 100 and the S&P 500. Future research will target enhanced sentiment analysis and broader market applications.

Keywords: stock forecasting; sentiment analysis; hybrid forecasting models; deep learning; machine learning; FinBERT.

UNCORRECTED PROOF - JRSN

1 Introduction

Forecasting stock market behaviour remains one of the most challenging problems in financial economics because of the market's complexity, volatility and sensitivity to external stimuli (Vuong et al., 2024; Li et al., 2024; Niu et al., 2022). The stock market is a critical component of the global economy, and this study focuses on three leading indices whose movements guide investment and risk-management decisions: the Standard & Poor's 500 (S&P 500), which tracks 500 large-capitalisation companies listed in the United States; the Financial Times Stock Exchange 100 (FTSE 100), which comprises the 100 largest UK-listed firms; and the Deutscher Aktienindex (DAX), which comprises 40 German companies traded on the Frankfurt Stock Exchange and serves as a core proxy for Europe. Accurate forecasts of these indices are valuable because investors seek to maximise returns while managing risk, yet index prices are volatile and are influenced by a wide range of factors (Bustos & Pomares-Quimbaya, 2020; Liu et al., 2022; Rundo et al., 2019; Yu et al., 2023).

Traditional approaches such as the Autoregressive Integrated Moving Average (ARIMA) model rely mainly on historical prices and technical indicators and overlook newer data sources such as online opinion and investor sentiment (Mutinda & Langat, 2024; Liapis et al., 2023; Mutinda et al., 2025). Machine learning models can capture non-linear relationships that such models miss (Siami-Namini et al., 2018), and large forecasting competitions show that combining statistical and machine learning methods is particularly effective (Makridakis et al., 2020; Mutinda & Geletu, 2025). At the same time, investor behaviour is increasingly shaped by real-time news and public sentiment on platforms such as Twitter, financial news sites and blogs (Cevik et al., 2022; Maqsood et al., 2020). Negative news can trigger panic selling and optimistic hype can inflate prices, movements that price-only models fail to anticipate.

The efficient market hypothesis holds that asset prices fully reflect all available information, so that the market cannot be beaten consistently. Behavioural finance qualifies this view: investor sentiment and psychology can drive prices away from fundamental values (Guo & Xie, 2024; Ghosh et al., 2024; Gao et al., 2020), and sentiment is correlated with both returns and volatility (Zhang et al., 2018). Advances in natural language processing (NLP) make it possible to measure sentiment in real time by classifying the tone of text such as news headlines as positive, negative or neutral (Mao et al., 2024). Translating unstructured, noisy text into numerical inputs is nevertheless difficult. Lexicon-based tools such as VADER and TextBlob assign polarity scores to words but struggle with context-dependent expressions such as sarcasm (Mao et al., 2024). Deep learning models such as bidirectional LSTMs process text in both directions and produce more informative sentiment representations (Islam et al., 2024; Sivadasan et al., 2024), and transformer models such as BERT (Devlin et al., 2019) and its financial adaptation FinBERT (Araci, 2019) capture nuanced context that lexicon methods miss.

Hybrid models that integrate sentiment with deep learning combine qualitative sentiment signals with quantitative market indicators and address limitations of conventional techniques (Jing et al., 2021; Lin et al., 2022). Sequence models trained on such combined inputs have been reported to improve directional accuracy and to reduce mean squared error by 7–15% relative to models that use historical prices only (Dahal et al., 2023; Guo & Xie, 2024). Key challenges remain: sourcing reliable sentiment data, measuring sentiment accurately, designing architectures that integrate sentiment with market data, and establishing whether the integrated approach outperforms sentiment-agnostic models.

This study addresses these challenges with a hybrid forecasting framework that integrates technical indicators (moving averages, Bollinger Bands and the Relative Strength Index) with sentiment extracted from financial news headlines by FinBERT. Each trading day is represented by a feature vector $X_t = (\text{Price}_t, \text{MA}_t, \text{BB}_t, \text{RSI}_t, \text{Sentiment}_t)$, so that the models learn joint patterns between market structure and investor sentiment. Deep learning models (LSTM, BiLSTM, CNN and DNN) are combined with Random Forest, and the hybrids are compared with their stand-alone counterparts.

1.1 Objectives and research questions

The specific objectives of the study are:

1. to collect and pre-process historical price and sentiment data for the S&P 500, FTSE 100 and DAX;
2. to engineer technical indicators (moving averages, Bollinger Bands, RSI) and extract sentiment features using FinBERT;
3. to develop hybrid forecasting models that combine deep learning architectures (LSTM, BiLSTM, CNN and DNN) with Random Forest;
4. to train and test these models under different configurations and data sets; and
5. to assess their performance using RMSE, MAE, R^2 and directional accuracy.

Correspondingly, the study addresses the following research questions:

RQ1. How can historical price and sentiment data be collected and processed for the three indices while ensuring data quality and compatibility?
RQ2. What is the effect of integrating technical indicators and FinBERT-extracted sentiment features on predictive performance?
RQ3. How can hybrid models that combine deep learning with Random Forest be designed for stock index prediction?
RQ4. How do different configurations affect training and testing outcomes?
RQ5. How do the hybrid models compare with their stand-alone counterparts?

1.2 Contributions

The contributions of this research are fivefold.

1. **Hybrid sentiment–technical framework.** We propose a unified framework that integrates FinBERT-extracted sentiment with a set of technical indicators and combines deep learning with Random Forest, explicitly modelling the joint influence of market structure and investor sentiment on index prices.
2. **Multi-index, cross-market validation.** Unlike most studies, which focus on a single market, we evaluate the approach on three major indices (S&P 500, FTSE 100 and DAX) in different regulatory

UNCORRECTED PROOF-JHSN

and economic environments.

3. **Systematic comparison of hybrid architectures.** Four deep learning models combined with Random Forest are benchmarked against five stand-alone models, each with and without sentiment.
4. **Comprehensive evaluation.** In addition to RMSE, MAE and R^2 we report directional accuracy, which is directly relevant to investment decisions; the use of FinBERT addresses the limitations of lexicon-based sentiment tools.
5. **Insights for investors and risk managers.** The results show when sentiment-enhanced models improve point forecasts and directional predictions, and when they do not.

The remainder of the paper is organised as follows. Section 2 reviews the literature, Section 3 describes the data and methodology, Section 4 presents and discusses the results, and Section 5 concludes.

2 Literature Review

Using textual sentiment from news and social media to improve stock market forecasting has gained considerable popularity. In an effort to quantify the effect of public opinion on financial markets, researchers have progressed from simple sentiment measures to complex machine learning and deep learning techniques. This section reviews the development of sentiment analysis in finance, its use in forecasting models, and the research gaps that motivate this study.

2.1 Evolution of financial sentiment analysis

Early approaches to financial sentiment analysis were lexicon-based and used predefined dictionaries of positive and negative words. Researchers built specialised lexicons, such as the Harvard IV or Loughran–McDonald dictionaries, to measure market sentiment (Chandra & Mondal, 2025; Liapis et al., 2023). These methods assign sentiment scores to text but often fail to recognise nuances such as sarcasm and context. As data volumes grew, machine learning classifiers such as Naïve Bayes and support vector machines were trained on labelled financial text and improved accuracy by learning from examples rather than fixed word lists (Liapis et al., 2023). For example, one study predicted Amazon, Apple and Microsoft stock prices using VADER sentiment from tweets and news headlines combined with XGBoost and reported an accuracy of about 89.8% (Chandra & Mondal, 2025), illustrating the value of combining lexicon-based features with machine learning.

Deep learning took over in the late 2010s by extracting complex text features automatically (Ji et al., 2021). Recurrent and convolutional neural networks were applied successfully to financial text (Islam et al., 2024); a comparative study found that a CNN combining price and news sentiment outperformed models using either input alone (Usmani & Shamsi, 2021). Recurrent models such as LSTM and GRU became popular for sequences of text and prices (Munyao et al., 2025; Mutinda & Yong, 2025a). A two-stream GRU architecture outperformed standard GRU and LSTM models in stock trend prediction with an accuracy of 66.32%, and its sentiment-tailored Stock2Vec embedding, based on the Harvard IV-4 lexicon, outperformed generic embeddings such as GloVe and Word2Vec (Minh et al., 2018).

The most recent stage is represented by transformer language models pre-trained on large corpora, such as BERT and RoBERTa, which can be fine-tuned for financial sentiment. One study fine-tuned RoBERTa on labelled stock tweets and substantially improved sentiment classification before feeding the scores into a prediction model (Guo & Xie, 2024). FinBERT, a BERT model adapted to financial text, improves markedly on earlier methods in financial sentiment classification (Araci, 2019). Hybrid architectures with multiple recurrent layers, attention mechanisms or capsule networks have also become popular; one review reports accuracies of up to 98% for capsule-network-enhanced RNNs on general benchmarks and lists the wide range of deep models now in use (Islam et al., 2024). In short, the field has moved from lexicon indices through traditional classifiers to deep learning and transformer models, each generation improving the measurement of investor sentiment.

2.2 Forecasting models with sentiment

Stock price forecasting has developed in parallel, from statistical models to hybrid deep learning frameworks. Traditional time-series models such as ARIMA handle trends and seasonality well but struggle with non-linear patterns and external inputs such as news sentiment. Machine learning models such as

UNCORRECTED PROOF-JHSN

random forests, support vector machines and neural networks capture non-linear relationships without strong assumptions (Li et al., 2025; Dong & Zhou, 2024). One study linked news-article trends to stock movements with random forests and SVMs and found better predictive power than statistical baselines, with sentiment indices as additional features (Liapis et al., 2023). Another study used econometric modelling and found that negative investor sentiment significantly raises market volatility while positive sentiment lowers it across all quantiles, which justifies the inclusion of sentiment in forecasting models (Cevik et al., 2022).

As deep learning gained popularity, researchers developed integrated models that combine price prediction with textual sentiment, typically in two components: one for sentiment extraction and one for price prediction. One study proposed a two-stage architecture in which an RNN forecasts price trends using sentiment from a fine-tuned RoBERTa model applied to social media (Guo & Xie, 2024). Many studies report that adding news or tweet sentiment improves accuracy relative to price data alone: an LSTM performed much better when headlines were included alongside stock features (Dahal et al., 2023), and combining historical prices with Twitter sentiment significantly improved predictions for Apple, Google and Tesla (Ouf et al., 2024). Hybrids that combine LSTM networks with random forests have also been proposed (Park et al., 2022; Ghosh et al., 2022).

Recent studies have reinforced the value of diverse data sources. Financial news from outlets such as Reuters, Bloomberg and The Wall Street Journal carries strong sentiment signals; Twitter and StockTwits posts have been combined to build richer sentiment profiles (Minh et al., 2018); and crowd-sourced platforms such as Reddit and Estimize have been integrated to fuse professional and retail sentiment. Technical indicators such as moving averages, Bollinger Bands, RSI and trading volume are commonly derived from prices to complement sentiment signals, and combined feature vectors fed to deep learning models such as LSTM substantially improve predictive accuracy.

2.3 Research gaps

Despite this progress, key gaps remain. First, econometric models such as ARIMA and GARCH rely on historical price series and do not account for external shocks such as geopolitical events, natural disasters or sudden shifts in investor mood (Vuong et al., 2024), so their predictions can be biased in the face of such disturbances. Second, although hybrid machine learning–deep learning architectures that blend, for instance, random forests with support vector machines or RNNs with transformer-derived sentiment have shown promise (Liapis et al., 2023; Guo & Xie, 2024), it remains unclear which combinations are most effective and how robust they are across markets. Third, sentiment analysis, even with transformers, is rarely optimised for real-time use because of the latency of processing high-volume news streams (Islam et al., 2024). Fourth, raw financial data are noisy, incomplete and punctuated by anomalies, and the literature offers only preliminary solutions for cleaning and imputing such inputs while preserving the patterns to be learnt (Cevik et al., 2022).

This study addresses these gaps by combining FinBERT-based sentiment with technical indicators, by systematically comparing four deep learning–Random Forest hybrids with their stand-alone counterparts across three major indices, and by applying a transparent pre-processing and imputation pipeline that preserves the continuity of the time series.

2.4 Positioning and novelty of this study

Hybrid frameworks that combine sentiment with deep learning already exist (Jing et al., 2021; Dahal et al., 2023; Liapis et al., 2023), and LSTM networks have been combined with random forests in a multi-task framework (Park et al., 2022) and compared with them for directional intraday forecasting (Ghosh et al., 2022). The present study differs from this work in four respects, summarised in Table 1. First, the Random Forest is used as an explicit second-stage residual corrector of the deep learning forecast (Section 3.5) rather than as a parallel model or a shared-representation learner. Second, the same hybrid construction is applied to four different network families (LSTM, BiLSTM, CNN and DNN), so that the effect of the hybrid design can be separated from that of the network architecture. Third, every model is estimated both with and without FinBERT sentiment, which isolates the contribution of sentiment for each architecture and each market, including the cases in which it is harmful. Fourth, the evaluation covers

UNCORRECTED PROOF-JFISM

three major indices from different regulatory and economic environments rather than a single market or a set of individual stocks.

Table 1: Design features of this study compared with common practice in sentiment-augmented hybrid stock forecasting.

Aspect	Common practice	This study
Combination of DL and RF	Parallel models, shared representation, or not specified	Random Forest as explicit residual corrector of the DL forecast (Equations 19–21, Algorithm 1)
Network families	One architecture, usually LSTM	Four architectures (LSTM, BiLSTM, CNN, DNN) with the same hybrid construction
Sentiment measurement	Lexicon tools (VADER, TextBlob)	FinBERT probabilities, $p_{\text{pos}} - p_{\text{neg}}$, aggregated daily
Contribution of sentiment	Reported for the best model only	With/without comparison for all nine models and three indices (Table 8)
Markets	Single index or individual stocks	S&P 500, FTSE 100 and DAX
Missing data	Row deletion	Three-step imputation preserving temporal continuity

3 Data and Methodology

Figure 1 summarises the workflow of the study, from data collection and pre-processing to the hybrid models and their evaluation. The components are described in turn below.

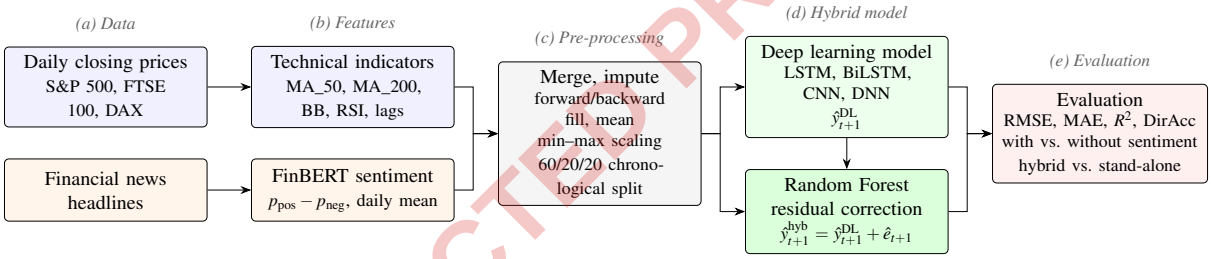


Figure 1: Workflow of the hybrid stock price forecasting framework: data collection, pre-processing and feature construction, deep learning forecast, Random Forest refinement and evaluation.

3.1 Data

The study uses daily data for three major stock indices, the S&P 500 (United States), the FTSE 100 (United Kingdom) and the DAX (Germany), from 1 January 2020 to 13 June 2023. This period captures recent market dynamics, including the volatility induced by the COVID-19 pandemic. All data are obtained from Yahoo Finance through the `yfinance` Python library. Two complementary types of data are collected: (i) daily closing prices of each index and (ii) the corresponding financial news headlines from Yahoo Finance, which are used to compute sentiment scores. After aligning prices and sentiment by date, the integrated data set contains 3,780 daily observations, each with nine fields: date, index ticker, closing price, MA_50, MA_200, BB_Upper, BB_Lower, RSI and sentiment score.

3.2 Pre-processing and feature engineering

Technical indicators. From the closing prices we derive indicators widely used in stock analysis: simple moving averages over 50 and 200 days (MA_50, MA_200), Bollinger Bands and the Relative Strength Index (RSI). Bollinger Bands lie two standard deviations above and below a 20-day simple moving average,

$$\text{BB}_{\text{upper}} = \text{MA}_{20} + 2\sigma_{20}, \tag{1}$$

$$\text{BB}_{\text{lower}} = \text{MA}_{20} - 2\sigma_{20}, \tag{2}$$

UNCORRECTED PROOF-JIGN

where σ_{20} is the standard deviation of the closing price over the same 20-day window. The RSI is a momentum oscillator between 0 and 100,

$$\text{RSI} = 100 - \frac{100}{1 + \text{RS}}, \tag{3}$$

$$\text{RS} = \frac{\text{average gain over 14 days}}{\text{average loss over 14 days}}. \tag{4}$$

Lagged features, price change and target. To capture short-term temporal dependencies, lags $t-1$, $t-2$ and $t-3$ are created for each numeric feature other than the price itself; three lags balance relevant history against redundancy, a common choice in time-series forecasting (Fischer & Krauss, 2018). A percentage price-change feature measures daily movements within each index,

$$\text{Price_Change}_t = \frac{P_t - P_{t-1}}{P_{t-1}} \times 100. \tag{5}$$

The target variable is the next day's closing price, obtained by shifting the price series one day forward, which gives one-step-ahead forecasts. Rows with missing values created by the lags are removed (Lu et al., 2020), and features are selected using a correlation threshold of 0.1.

Sentiment scores. News headlines are cleaned by removing special characters and stop-words, converting text to lower case and handling hashtags and mentions. The cleaned headlines are tokenised in the BERT input format and passed through FinBERT, a transformer language model adapted to the financial domain (Araci, 2019), which returns probabilities p_{pos}, p_{neu} and p_{neg} for positive, neutral and negative tone. The continuous sentiment score of a headline is

$$\text{Sentiment Score} = p_{\text{pos}} - p_{\text{neg}} \in [-1, +1], \tag{6}$$

from -1 (strongly negative) to $+1$ (strongly positive). Headlines are time-stamped and aligned with their trading day, and the scores of days with several headlines are averaged, giving one daily sentiment score per index.

Missing values. Merging market and sentiment data and computing indicators introduces missing values. Deleting the affected rows would break the temporal sequence, so a three-step imputation pipeline is applied:

1. **Forward fill** (last observation carried forward): recent conditions are assumed to persist in the absence of new data, which is reasonable for financial and sentiment series that change gradually.
2. **Backward fill** (next observation carried backward): fills gaps at the beginning of sequences or between forward-fill gaps, to maintain sequence continuity in the historical data.
3. **Mean imputation**: any remaining missing entry is replaced by the column mean, used only when no adjacent data are available.

Feature vector, scaling and data split. Each trading day is represented by the feature vector

$$X_t = \big(\text{Price}_t,\ \text{MA}_{50,t},\ \text{MA}_{200,t},\ \text{BB}_{\text{upper},t},\ \text{BB}_{\text{lower},t},\ \text{RSI}_t,\ \text{Sentiment}_t\big), \tag{7}$$

together with the lagged features described above. Because the features have different scales, they are normalised to $[0, 1]$ with min–max scaling,

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \tag{8}$$

The data are split chronologically into training (60%), validation (20%) and test (20%) sets. The earliest data form the training set, the middle segment is used for validation and hyperparameter tuning, and the most recent segment is the test set, which mimics a realistic forecasting situation.

UNCORRECTED PROOF-JHSN

3.3 Forecasting framework

The forecasting framework integrates three components: machine learning (Random Forest), which learns relationships between engineered features and future prices; deep learning (LSTM, BiLSTM, CNN and DNN), which captures temporal dependencies and local patterns in the price and indicator sequences; and sentiment analysis, which converts news headlines into daily sentiment scores. The forecast of the next-day closing price is

$$\hat{y}_{t+1} = f(X_t), \tag{9}$$

where $f(\cdot)$ is the learned forecasting function, realised either by a single model or by a hybrid. Every model is trained in two configurations, with technical indicators only (*without sentiment*) and with technical indicators plus sentiment (*with sentiment*), so that the contribution of sentiment can be assessed directly.

3.4 Stand-alone models

Random Forest. Random Forest (Breiman, 2001) is an ensemble of N decision trees, each trained on a bootstrap sample of the data with a random subset of features considered at each split, which decorrelates the trees. With $f_i(X_t)$ the prediction of the i-th tree, the forecast is the average

$$\hat{y}_t = \frac{1}{N}\sum_{i=1}^{N} f_i(X_t). \tag{10}$$

Long Short-Term Memory (LSTM). LSTM networks are recurrent networks that learn long-term dependencies through gating mechanisms, which mitigate the vanishing-gradient problem (Hochreiter & Schmidhuber, 1997; Mutinda & Yong, 2025b). An LSTM cell (Figure 2) updates its cell state c_t and hidden state h_t as

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \tag{11}$$
$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{12}$$
$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \tag{13}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c x_t + U_c h_{t-1} + b_c), \tag{14}$$
$$h_t = o_t \odot \tanh(c_t), \tag{15}$$

where i_t, f_t and o_t are the input, forget and output gates, x_t is the input at time t, σ is the logistic function, $\odot$ is the element-wise product, and W, U and b are learnable weights and biases. Our LSTM receives input sequences of 10 days and has two layers (64 units returning the full sequence, then 32 units), each followed by dropout of 0.3, and a single output neuron that predicts the scaled next-day price (Figure 3a).

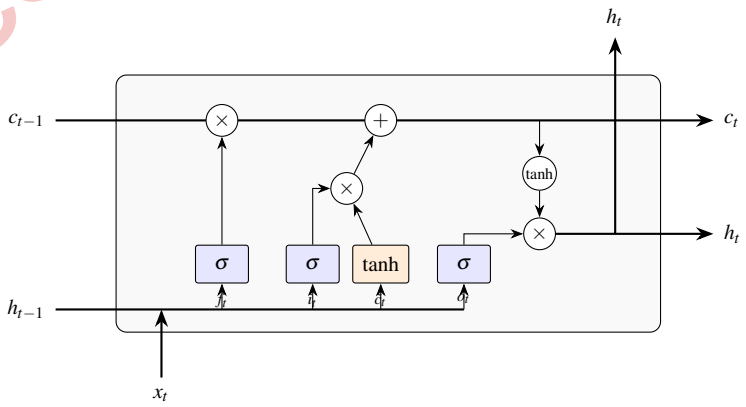


Figure 2: Structure of an LSTM cell. The forget gate f_t, input gate i_t and output gate o_t control the update of the cell state c_t and hidden state h_t (Equations 11–15).

UNCORRECTED PROOF-JHSN

Bidirectional LSTM (BiLSTM). A BiLSTM processes the input window in both the forward and backward directions and concatenates the two hidden states,

$$h_t^{\text{bi}} = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right], \tag{16}$$

which gives a fuller representation of the temporal dynamics within the window. Our BiLSTM uses the same window of 10 days, two bidirectional layers (64 and 32 units), dropout of 0.3 after each layer and a single output neuron (Figure 3b).

Convolutional Neural Network (CNN). One-dimensional convolutions extract local temporal patterns from time series:

$$y_t = \sum_{i=0}^{k-1} w_i x_{t+i} + b, \tag{17}$$

where w_i are the kernel weights and k is the kernel size. Our 1-D CNN has two convolutional layers (64 and 32 filters, kernel size 3, ReLU), each followed by max-pooling of size 2, a dense layer of 64 ReLU neurons with dropout of 0.3, and a single output neuron (Figure 3c). This architecture extracts short-term local features from financial time series (Mutinda & Yong, 2026).

Deep Neural Network (DNN). A DNN is a feed-forward network with several hidden layers,

$$h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right), \tag{18}$$

where $W^{(l)}$ and $b^{(l)}$ are the weights and biases of layer l and σ is the ReLU activation. Our DNN takes the scaled feature vector directly, without sequence construction, and has three hidden layers of 128, 64 and 32 neurons, each followed by dropout of 0.3, and a single output neuron (Figure 3d). It serves as a non-temporal baseline that shows the benefit of sequential modelling.

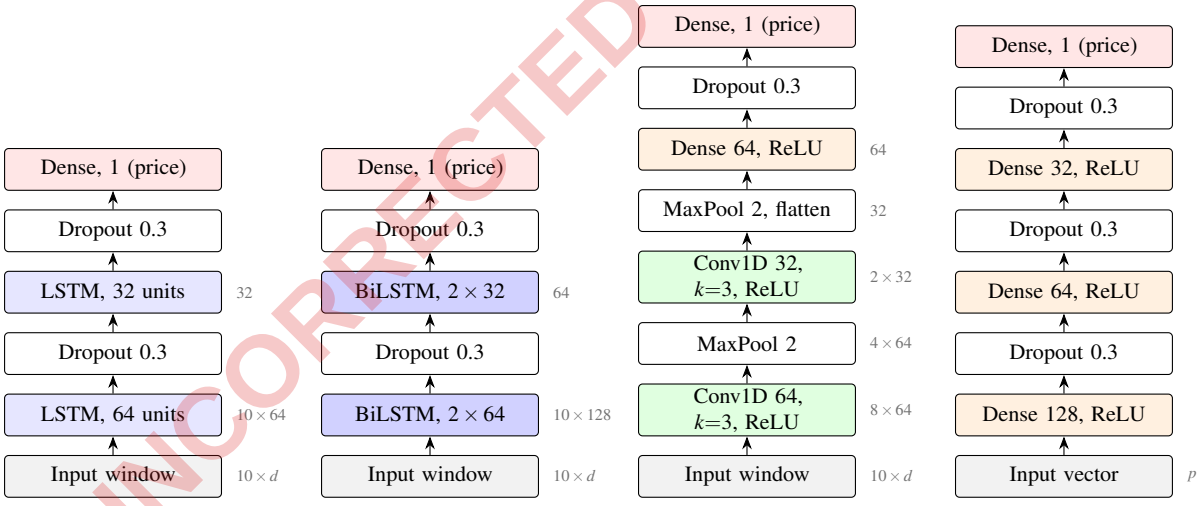


Figure 3: Architectures of the four deep learning models. Grey labels beside the boxes give the dimensions of the data after each stage; the input of the sequence models is a window of 10 days.

3.5 Hybrid deep learning–Random Forest models

In the hybrid models, the forecast of a deep learning model is refined by a Random Forest. The deep learning component captures temporal dependencies in the price, indicator and sentiment sequences, and the Random Forest corrects its residual errors by modelling non-linear patterns that the network misses (Park et al., 2022). Let $g(\cdot)$ denote a trained deep learning model and $\mathbf{X}_t = (X_{t-9}, \ldots, X_t)$ the input window. The procedure has three steps.

UNCORRECTED PROOF-JHSN

1. **Deep learning forecast.** The network is trained on the training set and produces

$$\hat{y}_{t+1}^{\mathrm{DL}} = g(\mathbf{X}_t). \tag{19}$$

2. **Residual learning.** The residuals of the network,

$$e_{t+1} = y_{t+1} - \hat{y}_{t+1}^{\mathrm{DL}}, \tag{20}$$

   are modelled by a Random Forest $h(\cdot)$ that uses the features and the network's forecast as inputs, $\hat{e}_{t+1} = h\big(X_t, \hat{y}_{t+1}^{\mathrm{DL}}\big)$.
3. **Hybrid forecast.** The final forecast adds the predicted residual to the network forecast,

$$\hat{y}_{t+1}^{\mathrm{hyb}} = \hat{y}_{t+1}^{\mathrm{DL}} + h\big(X_t, \hat{y}_{t+1}^{\mathrm{DL}}\big). \tag{21}$$

Algorithm 1 summarises the procedure. Four hybrids are built in this way: LSTM + RF, BiLSTM + RF, CNN + RF and DNN + RF.

Algorithm 1 Hybrid deep learning + Random Forest forecasting

Require: scaled features $\{X_t\}$, targets $\{y_{t+1}\}$, chronological training, validation and test sets, window length $L = 10$

Build input windows $\mathbf{X}_t = (X_{t-L+1}, \ldots, X_t)$ for the sequence models
Train the deep learning model g on the training set (Adam, MSE loss, early stopping on the validation set)
Compute $\hat{y}_{t+1}^{\mathrm{DL}} = g(\mathbf{X}_t)$ and residuals $e_{t+1} = y_{t+1} - \hat{y}_{t+1}^{\mathrm{DL}}$ on the training set
Tune and train the Random Forest h on $\{(X_t, \hat{y}_{t+1}^{\mathrm{DL}}), e_{t+1}\}$ (grid search, time-series cross-validation)
for each day t in the test set **do**
    $\hat{y}_{t+1}^{\mathrm{hyb}} \leftarrow g(\mathbf{X}_t) + h\big(X_t, g(\mathbf{X}_t)\big)$
end for
Evaluate $\hat{y}^{\mathrm{hyb}}$ with RMSE, MAE, R^2 and DirAcc (Section 3.6)

3.6 Evaluation metrics

Forecasting performance is evaluated on the training, validation and test sets of each index with four metrics, where N is the number of observations, y_i the actual and $\hat{y}_i$ the predicted closing price. The root mean squared error measures the typical size of errors in the units of the target and gives more weight to large errors,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{22}$$

The mean absolute error treats all deviations equally and complements the RMSE,

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{23}$$

The coefficient of determination gives the share of the variance of the target explained by the model; values close to one indicate a good fit, and negative values mean that the model performs worse than a constant forecast equal to the mean $\bar{y}$,

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{24}$$

Directional accuracy measures the share of correctly predicted directions of price movement, which is critical for trading decisions,

$$\mathrm{DirAcc} = \frac{1}{N-1}\sum_{i=1}^{N-1}\mathbb{I}\big(\mathrm{sign}(y_{i+1} - y_i) = \mathrm{sign}(\hat{y}_{i+1} - \hat{y}_i)\big), \tag{25}$$

UNCORRECTED PROOF JHSN

where $\mathbb{I}(\cdot)$ equals one if the condition holds and zero otherwise.

3.7 Hyperparameters

The Random Forest is tuned by grid search with five-fold time-series cross-validation, which preserves the chronological order of the observations in each fold and prevents look-ahead bias. Table 2 lists the search space. The combination with the lowest mean squared error averaged over the five folds is selected, separately for each index and for each configuration (with and without sentiment).

Table 2: Grid search space for the Random Forest hyperparameters (five-fold time-series cross-validation).

Parameter	Description	Values searched
Number of estimators	Number of trees in the forest	50, 100, 150, 200
Maximum depth	Maximum depth of each tree	5, 10, 15, none
Maximum features	Features considered at each split	sqrt, log2, 0.3, 0.5
Minimum samples split	Minimum samples to split an internal node	2, 5, 10, 15
Minimum samples leaf	Minimum samples at a leaf node	1, 2, 4, 8
Bootstrap	Whether bootstrap samples are used	True, False
Criterion	Split-quality function	squared error, absolute error, Friedman MSE, Poisson

The architectures of the deep learning models are fixed to keep the experiments comparable and to isolate the effect of sentiment. All networks are trained with the Adam optimiser (learning rate 0.001), mean squared error loss, a batch size of 32 and at most 100 epochs, with early stopping when the validation loss has not improved for 10 consecutive epochs. Table 3 lists the architecture-specific settings.

Table 3: Fixed hyperparameters of the deep learning models.

Setting	LSTM	BiLSTM	1-D CNN	DNN
Input	10-day window	10-day window	10-day window	feature vector
Layers	2 LSTM	2 BiLSTM (concatenated)	2 Conv1D + 2 max-pool	3 dense
Units / filters	64, 32	64, 32 per direction	64, 32 (kernel 3, pool 2)	128, 64, 32
Dense layer before output	–	–	64 units	–
Hidden activation	ReLU	ReLU	ReLU	ReLU
Dropout	0.3 after each layer	0.3 after each layer	0.3 before output	0.3 after each layer
Output layer	1 unit, linear	1 unit, linear	1 unit, linear	1 unit, linear

Common to all models: Adam (learning rate 0.001), MSE loss, batch size 32, at most 100 epochs, early stopping with patience 10 on the validation loss.

UNCORRECTED PROOF-JHSN

4 Results and Discussion

4.1 Descriptive statistics

The analysis uses daily data for the DAX, FTSE 100 and S&P 500 from 1 January 2020 to 13 June 2023, covering 1,260 days per index. The data set contains closing prices, sentiment scores and technical indicators (50- and 200-day moving averages, Bollinger Bands and RSI), together with their lags at $t-1$, $t-2$ and $t-3$. Table 4 reports descriptive statistics by index.

The S&P 500 shows the largest price dispersion (standard deviation 66.45), followed closely by the FTSE 100 (65.67), whereas the DAX shows more stable price patterns (21.00). The 50-day moving average tracks price dynamics closely, while the 200-day moving average is smoother and reflects long-term behaviour, notably for the S&P 500. RSI values are centred around 50 for all indices, so the markets were neither persistently overbought nor oversold during the sample period.

The histograms in Figure 4 guide model selection and feature engineering. The price distribution is positively skewed, with most observations between 100 and 200, indicating infrequent but possible extreme highs, a pattern typical of growing indices. The 50-day moving average has a similar shape, while the 200-day moving average is multimodal, which suggests regime shifts between bullish and bearish periods and calls for models that accommodate non-linear trends. Both Bollinger Bands are positively skewed, and the upper band is more widely spread, reflecting price expansions above the moving-average

Table 4: Summary statistics of prices, technical indicators and sentiment by index, January 2020 to June 2023.

Statistic	DAX	FTSE 100	S&P 500	Statistic	DAX	FTSE 100	S&P 500
Price				*RSI*			
Mean	123.24	204.06	183.46	Mean	50.75	52.17	52.74
Std. dev.	21.00	65.67	66.45	Std. dev.	12.19	11.95	10.94
Min.	80.79	86.90	91.08	Min.	15.08	16.84	24.77
25%	106.84	161.79	131.38	25%	42.96	44.73	44.80
Median	122.58	202.73	165.45	Median	50.32	52.77	52.98
75%	137.37	239.82	235.07	75%	59.49	60.18	60.25
Max.	176.74	376.01	326.06	Max.	81.61	90.87	82.87
MA_50				*BB_Upper*			
Mean	126.94	208.17	179.98	Mean	133.59	219.37	193.36
Std. dev.	26.87	64.54	65.08	Std. dev.	26.39	69.13	69.97
Min.	83.54	91.57	99.47	Min.	82.35	98.99	100.39
25%	108.43	164.45	128.31	25%	114.97	173.97	140.43
Median	123.21	204.21	161.65	Median	132.48	217.04	177.75
75%	139.53	249.61	233.36	75%	147.56	258.20	250.00
Max.	224.58	355.91	315.57	Max.	262.74	393.87	346.96
MA_200				*BB_Lower*			
Mean	135.44	226.01	168.46	Mean	116.58	192.30	170.40
Std. dev.	34.36	61.59	57.58	Std. dev.	23.49	61.68	61.99
Min.	90.45	95.91	98.99	Min.	73.05	89.28	90.38
25%	112.25	186.49	116.28	25%	99.07	153.31	121.61
Median	125.69	218.91	158.75	Median	115.31	189.73	153.17
75%	144.78	292.78	201.89	75%	128.43	226.26	214.36
Max.	216.28	317.94	303.23	Max.	223.82	349.67	306.49
				Sentiment			
				Mean	−0.020	−0.018	−0.020
				Std. dev.	0.570	0.572	0.570
				Min.	−1.000	−1.000	−1.000
				25%	−0.511	−0.511	−0.516
				Median	−0.019	−0.014	−0.015
				75%	0.477	0.477	0.466
				Max.	1.000	1.000	1.000

baseline. The RSI distribution is nearly symmetric and bell-shaped around 50, consistent with a bounded momentum oscillator. Sentiment, in contrast, is approximately uniformly distributed between -1 and $+1$, with no central tendency; it is a high-variance, behaviourally driven feature that complements the technical indicators in non-linear hybrid models.

Figure 5 compares the Bollinger Bands of the three indices on a common vertical axis. The wider bands of the S&P 500 and FTSE 100 relative to the DAX underline their higher volatility.

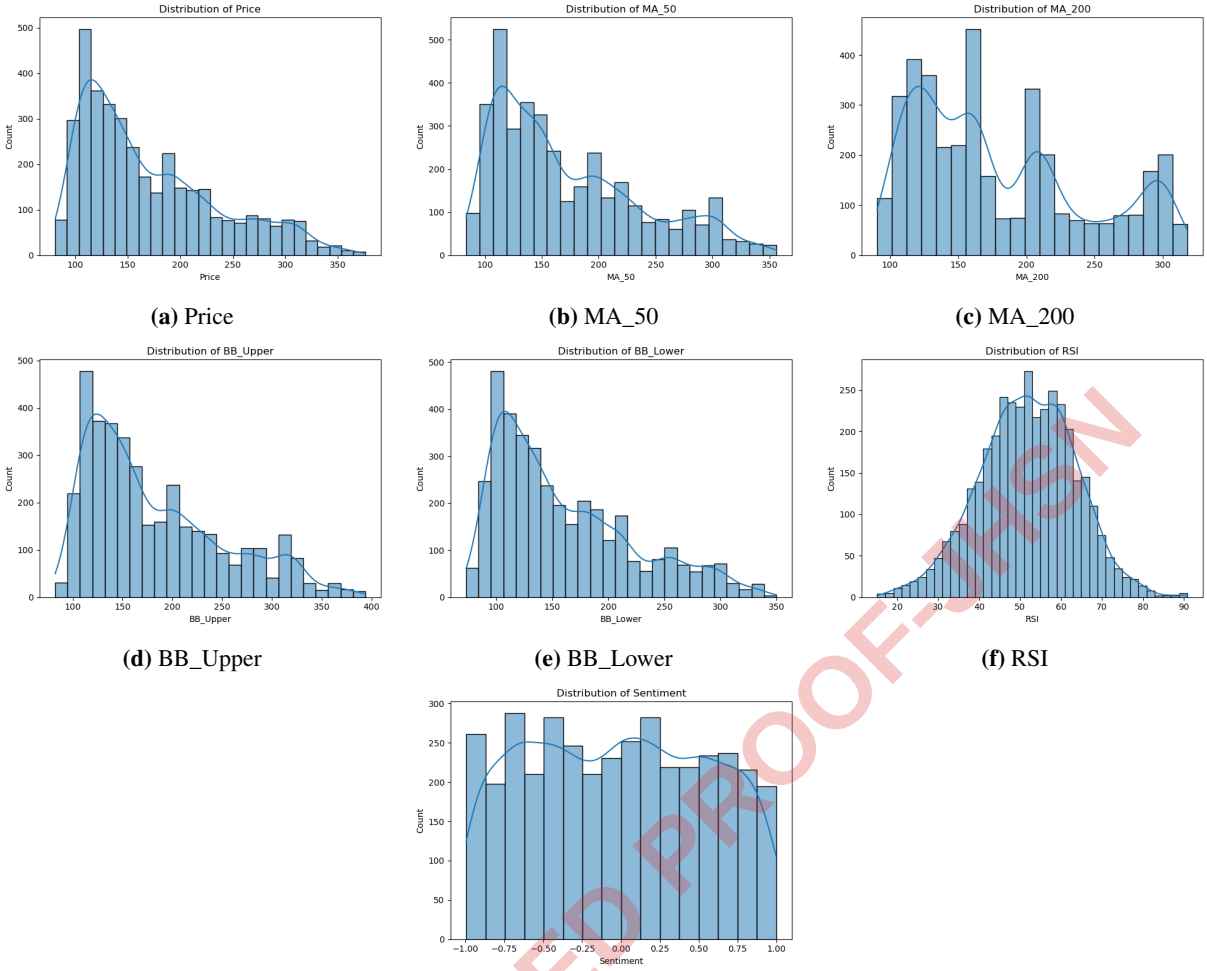
4.2 Performance of the hybrid models

Table 5 reports the test-set performance of the four hybrid models for each index, with and without sentiment.

DAX. LSTM + RF performs best on the DAX with sentiment, with an RMSE of 1.97, an MAE of 1.53, an R^2 of 0.98 and a DirAcc of 0.94, indicating high predictive accuracy and trend-prediction ability. Removing sentiment increases the RMSE to 2.14 and lowers the DirAcc to 0.90, which highlights the role of sentiment for the DAX. BiLSTM + RF, CNN + RF and DNN + RF have higher RMSE values with sentiment (1.98, 2.26 and 2.03) and lower directional accuracy (0.91, 0.90 and 0.88), suggesting that the LSTM + RF architecture is best suited to the DAX.

FTSE 100. Results for the FTSE 100 are mixed. LSTM + RF has the same RMSE (2.35) and DirAcc (0.88) in both modes, so sentiment has a negligible effect. BiLSTM + RF performs better without sentiment (RMSE 2.38 vs. 2.60), suggesting that sentiment introduces noise for this index. CNN + RF and DNN + RF reach RMSE values of 2.81 and 2.87 with sentiment and DirAcc values of 0.85 and 0.81,

UNCORRECTED PROOF - JHSN



(a) Price **(b)** MA_50 **(c)** MA_200

(d) BB_Upper **(e)** BB_Lower **(f)** RSI

(g) Sentiment

Figure 4: Distributions of prices, technical indicators and sentiment, pooled over the three indices.

UNCORRECTED PROOF-JHSN

reflecting difficulties in capturing the temporal dependencies of the FTSE 100.

S&P 500. Performance on the S&P 500 is similar across models, with RMSE values from 2.73 (DNN + RF with sentiment) to 2.81 (CNN + RF without sentiment) and DirAcc values between 0.87 and 0.89. Sentiment has a small effect; for example, LSTM + RF has a DirAcc of 0.89 in both modes, possibly because the index's diverse drivers dilute the predictive value of sentiment.

Across the three indices, LSTM + RF has the lowest RMSE for the DAX (1.97) and the FTSE 100 (2.35), and an RMSE within 0.01 of the best model for the S&P 500 (2.74 vs. 2.73 for DNN + RF), with $R^2 = 0.98$ for all three indices.

Figure 6 shows the LSTM + RF forecasts and the training and validation loss curves for each index. For the DAX (panels a and b), the model converges quickly without large divergence between training and validation loss, indicating good generalisation without obvious over-fitting, and the forecasts track the actual prices closely through both the upswing and the subsequent pull-back, with small deviations only at sharp turning points. For the FTSE 100 (panels c and d), the training loss falls rapidly over the first five epochs and then levels off, with small early fluctuations, and the forecasts follow the major rebounds and pull-backs with slight deviations at the sharpest turning points. For the S&P 500 (panels e and f), the training loss falls from 0.027 to about 0.008 while the validation loss drops from 0.13 to about 0.05 and then oscillates modestly; the close alignment of the two curves indicates convergence without severe over-fitting, and the forecasts follow the actual series through rebounds, peaks and pull-backs, with minor deviations at sharp local extrema.

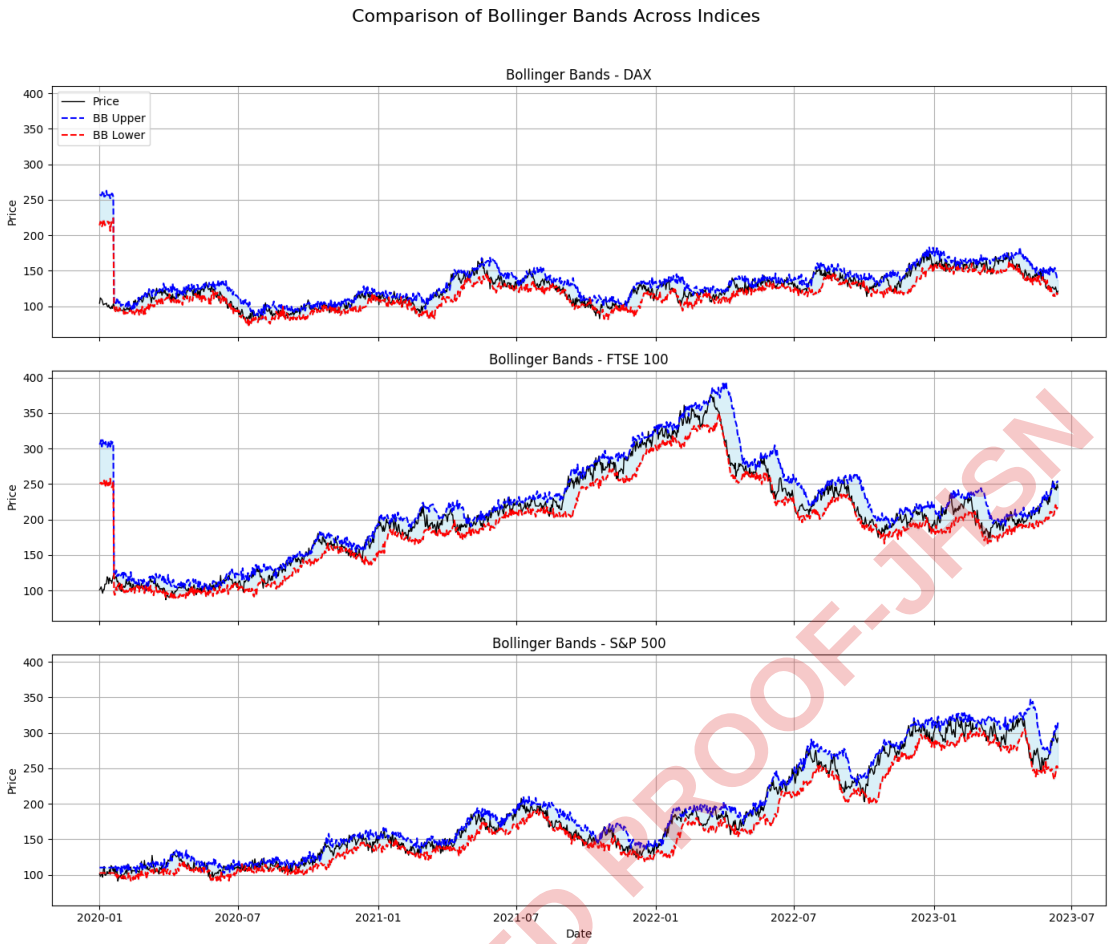


Figure 5: Bollinger Bands of the DAX, FTSE 100 and S&P 500, 2020–2023: price (black), upper and lower bands (blue and red dashed lines) and band width (shaded). The common vertical axis allows the band spread to be compared across indices.

UNCORRECTED PROOF-JHSN

4.3 Performance of the stand-alone models

Table 6 reports the results of the five stand-alone models, with and without sentiment. Lower RMSE and MAE and higher R^2 and DirAcc indicate better performance; a negative R^2 means that the model performs worse than a constant forecast equal to the mean.

DAX. The best stand-alone model for the DAX is the CNN with sentiment, with the lowest RMSE (14.15) and MAE (12.10) and an R^2 of -0.04, close to zero. It therefore barely outperforms a constant mean forecast, but it is far better than the Random Forest ($R^2 = -2.89$). The LSTM with sentiment also performs well (RMSE 14.59) and has the highest directional accuracy (0.56). Sentiment reduces the RMSE of the LSTM from 19.64 to 14.59 and of the CNN from 19.85 to 14.15; the improvement is modest for the BiLSTM and the DNN. The Random Forest performs worst, with a very high RMSE and a strongly negative R^2.

FTSE 100. The BiLSTM without sentiment has the lowest RMSE (13.12) and MAE (9.65) and one of only two positive R^2 values (0.35). Adding sentiment raises its RMSE to 18.97. Similarly, the LSTM without sentiment (RMSE 13.52, $R^2 = 0.31$) outperforms its sentiment-augmented version. For this index, sentiment features appear to introduce noise, possibly because of the market's specific characteristics or lower sensitivity to the news flow. The CNN with sentiment has the highest directional accuracy (0.55) despite a higher RMSE, suggesting reasonable trend prediction even when point errors are large.

S&P 500. The CNN with sentiment is clearly the best model, with an RMSE of 18.46, an MAE of 16.03 and an R^2 of -0.04, whereas the same model without sentiment has an RMSE of 31.09. The LSTM and BiLSTM also benefit from sentiment, but their RMSE values remain much higher (46.53 to 73.05) and

Table 5: Test-set performance of the hybrid models (LSTM + RF, BiLSTM + RF, CNN + RF and DNN + RF) with and without sentiment features.

Model	Index	Mode	RMSE	MAE	R^2	DirAcc
LSTM + RF	DAX	With sentiment	1.97	1.53	0.98	0.94
		Without sentiment	2.14	1.61	0.98	0.90
	FTSE 100	With sentiment	2.35	1.82	0.98	0.88
		Without sentiment	2.35	1.87	0.98	0.88
	S&P 500	With sentiment	2.74	2.21	0.98	0.89
		Without sentiment	2.77	2.17	0.98	0.89
BiLSTM + RF	DAX	With sentiment	1.98	1.58	0.98	0.91
		Without sentiment	2.01	1.57	0.98	0.90
	FTSE 100	With sentiment	2.60	2.03	0.97	0.87
		Without sentiment	2.38	1.88	0.98	0.88
	S&P 500	With sentiment	2.74	2.18	0.98	0.89
		Without sentiment	2.78	2.27	0.98	0.88
CNN + RF	DAX	With sentiment	2.26	1.68	0.97	0.90
		Without sentiment	2.34	1.76	0.97	0.89
	FTSE 100	With sentiment	2.81	2.27	0.97	0.85
		Without sentiment	2.46	1.95	0.98	0.86
	S&P 500	With sentiment	2.76	2.20	0.98	0.89
		Without sentiment	2.81	2.34	0.98	0.88
DNN + RF	DAX	With sentiment	2.03	1.59	0.98	0.88
		Without sentiment	2.05	1.58	0.98	0.85
	FTSE 100	With sentiment	2.87	2.25	0.97	0.81
		Without sentiment	2.51	1.96	0.98	0.84
	S&P 500	With sentiment	2.73	2.19	0.98	0.87
		Without sentiment	2.78	2.24	0.98	0.88

their R^2 values are strongly negative (-5.64 to -15.37). The Random Forest again performs very poorly, with an RMSE above 105 and an R^2 below -26, confirming that tree-based methods without temporal modelling are inadequate for forecasting price levels.

Overall, the recurrent networks (LSTM and BiLSTM) outperform the stand-alone Random Forest for every index and mode, and all four networks do so for the DAX and the S&P 500; for the FTSE 100, the CNN and the DNN have higher errors than the Random Forest. Directional accuracy remains modest, between 0.40 and 0.56, so predicting the direction of daily price changes remains difficult for the stand-alone models. These results provide the baseline against which the hybrid models are assessed.

4.4 Hybrid versus stand-alone models

Table 7 brings together, from Tables 5 and 6, the results of each deep learning model with sentiment and of its hybrid counterpart. The stand-alone models have large errors and negative R^2 values, whereas the hybrids reduce the RMSE by 84–96% for every architecture and index, attain R^2 values of 0.97–0.98 and directional accuracies of 0.81–0.94. The hybrids thus produce tighter point forecasts and also predict the direction of price movements more accurately (Ghosh et al., 2022).

The superior performance of the hybrids can be explained by the complementary strengths of their components. The deep learning models capture temporal dependencies in the price, indicator and sentiment sequences, while the Random Forest corrects the residuals by modelling non-linear patterns that the networks miss (Park et al., 2022). This complementarity is absent from the stand-alone models: the Random Forest alone does not model sequential structure, and the networks alone do not correct their residual errors.

4.5 Effect of sentiment

Table 8 summarises, for every model and index, the percentage change in RMSE when sentiment features are added (negative values mean that sentiment improves accuracy). The effect of sentiment differs clearly

UNCORRECTED PROOF JHSN

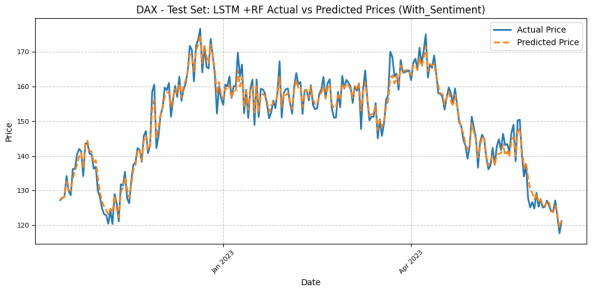
Table 6: Test-set performance of the stand-alone models with and without sentiment features.

Model	Index	Mode	RMSE	MAE	R^2	DirAcc
RF	DAX	With sentiment	28.56	25.16	−2.89	0.53
		Without sentiment	28.35	24.95	−2.83	0.53
	FTSE 100	With sentiment	19.19	16.22	−0.43	0.50
		Without sentiment	19.21	16.27	−0.43	0.46
	S&P 500	With sentiment	105.81	104.02	−26.43	0.54
		Without sentiment	105.01	103.20	−26.02	0.49
LSTM	DAX	With sentiment	14.59	12.97	−0.10	0.56
		Without sentiment	19.64	17.64	−1.00	0.54
	FTSE 100	With sentiment	18.13	13.74	−0.24	0.51
		Without sentiment	13.52	10.12	0.31	0.53
	S&P 500	With sentiment	63.96	62.49	−11.54	0.55
		Without sentiment	73.05	71.47	−15.37	0.51
BiLSTM	DAX	With sentiment	15.17	13.55	−0.19	0.53
		Without sentiment	15.95	14.02	−0.32	0.48
	FTSE 100	With sentiment	18.97	14.69	−0.36	0.40
		Without sentiment	13.12	9.65	0.35	0.47
	S&P 500	With sentiment	46.53	44.77	−5.64	0.52
		Without sentiment	62.20	60.57	−10.87	0.49
DNN	DAX	With sentiment	23.04	20.46	−1.53	0.55
		Without sentiment	25.35	22.91	−2.07	0.42
	FTSE 100	With sentiment	21.81	18.62	−0.85	0.46
		Without sentiment	21.85	19.02	−0.85	0.48
	S&P 500	With sentiment	37.11	35.69	−2.38	0.50
		Without sentiment	38.68	36.88	−2.67	0.48
CNN	DAX	With sentiment	14.15	12.10	−0.04	0.47
		Without sentiment	19.85	17.64	−1.04	0.45
	FTSE 100	With sentiment	24.31	21.20	−1.24	0.55
		Without sentiment	19.80	16.32	−0.48	0.48
	S&P 500	With sentiment	18.46	16.03	−0.04	0.53
		Without sentiment	31.09	29.23	−1.96	0.48

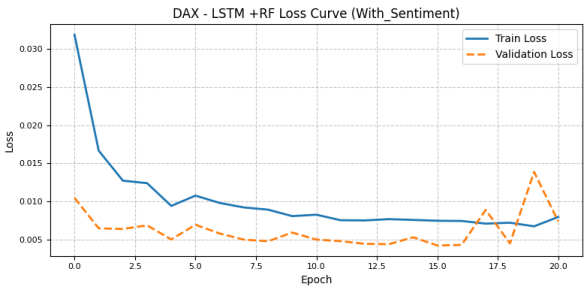
Table 7: RMSE and directional accuracy of each deep learning model and its hybrid counterpart (models with sentiment), compiled from Tables 5 and 6.

		Stand-alone		Hybrid (+RF)		
Model	Index	RMSE	DirAcc	RMSE	DirAcc	RMSE reduction (%)
LSTM	DAX	14.59	0.56	1.97	0.94	86.5
	FTSE 100	18.13	0.51	2.35	0.88	87.0
	S&P 500	63.96	0.55	2.74	0.89	95.7
BiLSTM	DAX	15.17	0.53	1.98	0.91	86.9
	FTSE 100	18.97	0.40	2.60	0.87	86.3
	S&P 500	46.53	0.52	2.74	0.89	94.1
CNN	DAX	14.15	0.47	2.26	0.90	84.0
	FTSE 100	24.31	0.55	2.81	0.85	88.4
	S&P 500	18.46	0.53	2.76	0.89	85.0
DNN	DAX	23.04	0.55	2.03	0.88	91.2
	FTSE 100	21.81	0.46	2.87	0.81	86.8
	S&P 500	37.11	0.50	2.73	0.87	92.6

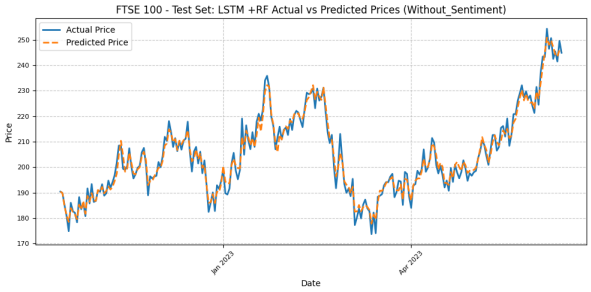
UNCORRECTED PROOF IJISN



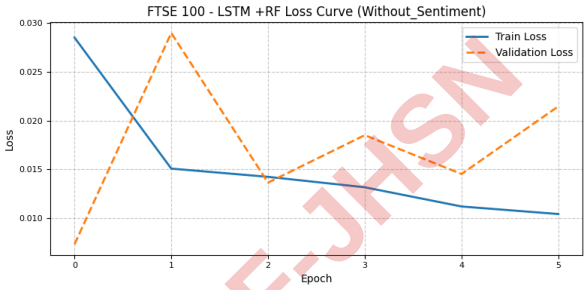
(a) DAX: actual vs. predicted prices



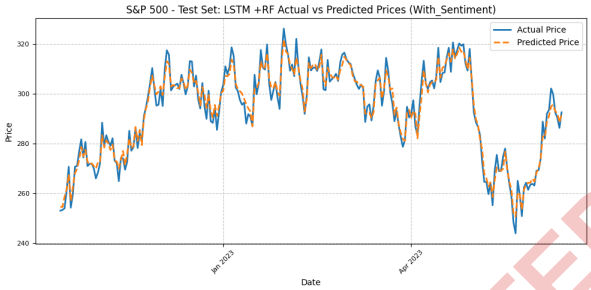
(b) DAX: training and validation loss



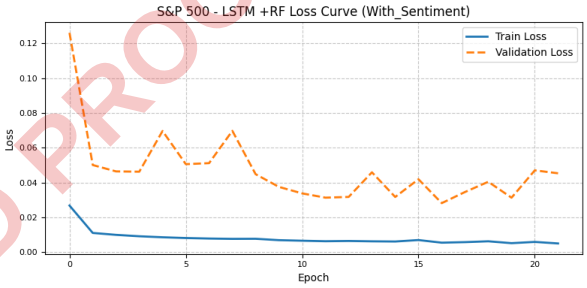
(c) FTSE 100: actual vs. predicted prices



(d) FTSE 100: training and validation loss



(e) S&P 500: actual vs. predicted prices



(f) S&P 500: training and validation loss

Figure 6: LSTM + RF with sentiment: actual vs. predicted test-set prices and training/validation loss curves for the DAX, FTSE 100 and S&P 500.

UNCORRECTED PROOF-JHSN

across indices. For the DAX, adding sentiment lowers the RMSE of all models except the stand-alone Random Forest, by 1.0–7.9% for the hybrids and by up to 28.7% for the stand-alone networks; the largest hybrid gain is for LSTM + RF, whose DirAcc also rises from 0.90 to 0.94. For the S&P 500, sentiment also lowers the RMSE of all models except the Random Forest, but the gains for the hybrids are small (1.1–1.8%), whereas those for the stand-alone networks are large (4.1–40.6%). For the FTSE 100, sentiment leaves the Random Forest, DNN and LSTM + RF essentially unchanged and increases the RMSE of the other six models, by 9.2–14.3% for three of the hybrids and by 22.8–44.6% for the stand-alone LSTM, BiLSTM and CNN.

These differences call for a critical interpretation rather than a general claim that sentiment improves forecasts. Several mechanisms are consistent with the pattern. First, news may be incorporated into prices quickly, especially for heavily covered markets, so that headline sentiment largely reflects same-day movements and adds little information about the next day; this would explain the small hybrid gains for the S&P 500. Second, the headlines were collected from a single English-language source, Yahoo Finance, whose coverage may match the FTSE 100's internationally exposed constituents, which are driven by commodity prices and exchange rates, less well than the other indices. Third, adding a noisy, high-variance feature (Figure 4g) increases the risk of over-fitting, especially for high-capacity networks such as the BiLSTM, which is consistent with the large deteriorations of the stand-alone networks on the FTSE 100. Finally, the effect of sentiment differs between architectures for the same index, which shows that its value depends on how it is combined with the other inputs. Within this sample, sentiment is most consistently useful for the DAX, and incorporating additional information such as macroeconomic

Table 8: Percentage change in RMSE when sentiment features are added, computed from Tables 5 and 6 (negative values: sentiment reduces the error).

Model	DAX	FTSE 100	S&P 500
RF	0.7	−0.1	0.8
LSTM	−25.7	34.1	−12.4
BiLSTM	−4.9	44.6	−25.2
CNN	−28.7	22.8	−40.6
DNN	−9.1	−0.2	−4.1
LSTM + RF	−7.9	0.0	−1.1
BiLSTM + RF	−1.5	9.2	−1.4
CNN + RF	−3.4	14.2	−1.8
DNN + RF	−1.0	14.3	−1.8

indicators may be needed to improve forecasts for the FTSE 100 and the S&P 500 (Mohan et al., 2019).

UNCORRECTED PROOF-JHSN

5 Conclusion

This study developed and evaluated hybrid models that combine deep learning networks with Random Forest for forecasting the S&P 500, FTSE 100 and DAX, using technical indicators and FinBERT-based sentiment features. The hybrid models consistently outperform their stand-alone counterparts, achieving lower errors and higher directional accuracy for all three indices. The two components are complementary: the deep learning models capture complex temporal dependencies, while the Random Forest corrects residual non-linear patterns that the networks miss. This combination produces more accurate forecasts than either class of model alone.

Sentiment features generally improve forecasting performance, but the effect depends on the market and the model. Sentiment is most consistently useful for the DAX, where it lowers the error of every deep learning and hybrid model. For the S&P 500 its benefit for the hybrids is small, and for the FTSE 100 it leaves some models unchanged and degrades others. The value of sentiment therefore depends on the underlying market dynamics and on how it is combined with other inputs, and a one-size-fits-all approach to feature integration is unlikely to be optimal.

The proposed architecture, which combines deep learning for temporal modelling with an ensemble method for residual correction, can be applied to other financial assets, including individual equities, cryptocurrencies such as Bitcoin and Ethereum, commodity futures such as gold, oil and agricultural products, and foreign exchange rates. The same principles of combining several data sources (technical indicators, sentiment and possibly macroeconomic variables) and blending different model families apply to these asset classes, and the hybrid approach is particularly valuable when the series are non-linear, non-stationary and subject to regime shifts.

5.1 Limitations

Several limitations must be acknowledged. The analysis covers a limited period and three indices, which may limit the generalisability of the quantitative findings to other markets or extreme regimes. Although regularisation and early stopping were applied, machine learning models may still fit noise in the training data, especially when the signal-to-noise ratio is low. The sentiment features depend on the accuracy and timeliness of the classification of news headlines from a single source, and noisy, biased or misaligned sentiment data can reduce or even reverse their predictive benefit. The evaluation is based on next-day price levels, and the features include current and lagged price levels; because tomorrow's price is close to today's, price-level metrics such as R^2 can be high even for forecasts with little ability to predict price changes, and min–max scaling and backward filling over the full sample can introduce look-ahead information. The large accuracy gains of the hybrids over the stand-alone models (Table 7) should therefore be confirmed with return-based targets, scaling and imputation fitted on the training data only, and comparison with a random-walk benchmark. The study also does not include econometric benchmarks (ARIMA, GARCH), transformer-based forecasters or formal tests of forecast accuracy. Finally, real-time deployment faces practical challenges, including the latency of processing high-volume text streams and the need for continual retraining as market conditions change.

5.2 Future research

Future research should address these limitations by extending the data to longer horizons and to other asset classes, including cryptocurrencies, commodities and exchange rates. The hybrid framework should be benchmarked against econometric models such as ARIMA and GARCH and against transformer-based forecasters such as the Temporal Fusion Transformer and Informer, and differences in accuracy should be assessed with formal tests such as the Diebold–Mariano test. Attention mechanisms could help the models focus on the most relevant time windows and news events, and explainable artificial intelligence methods would reveal which features and temporal patterns drive the predictions, which is important for adoption by traders and risk managers. Adaptive retraining and online learning could maintain forecast accuracy as market conditions evolve. By addressing these points, the approach can develop into a robust, interpretable and practically useful tool for market participants.

UNCORRECTED PROOF-JHSY

Author contributions

Tiapo Lydia Chimbaira: Conceptualization, Methodology, Writing – original draft, Visualization.
Amos Kipkorir Langat: Supervision, Project administration, Writing – review & editing, Validation.
Ranya Hamid Elnour Salim: Formal analysis, Validation, Writing – review & editing.
Mirriam Chepleting Sitienei: Data curation, Investigation, Writing – review & editing.
John Kamwele Mutinda: Software, Visualization, Data curation, Writing – review & editing.
David Chepkonga: Supervision, Project administration, Writing – review & editing, Validation.n.

Funding

No funding was received for this research.

Data availability

The data that support the findings of this study are available on request from the corresponding author.

Code availability

The code used for data analysis and visualisation in this study is available on request from the corresponding author.

Declarations

Ethical approval: This study did not involve humans or animals, so ethical approval was not required. **Competing interests:** The authors declare no competing interests. **Consent to participate:** Not applicable. **Consent for publication:** Not applicable.

UNCORRECTED PROOF-JHSN

References

Araci, D. (2019). *FinBERT: Financial sentiment analysis with pre-trained language models* (arXiv:1908.10063). arXiv. https://doi.org/10.48550/arXiv.1908.10063

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

Bustos, O., & Pomares-Quimbaya, A. (2020). Stock market movement forecast: A systematic review. *Expert Systems with Applications*, *156*, Article 113464. https://doi.org/10.1016/j.eswa.2020.113464

Cevik, E., Kirci Altinkeski, B., Cevik, E. I., & Dibooglu, S. (2022). Investor sentiments and stock markets during the COVID-19 pandemic. *Financial Innovation*, *8*(1), Article 69. https://doi.org/10.1186/s40854-022-00375-0

Chandra, J., & Mondal, A. C. (2025). Studies of sentiment analysis for stock market prediction using machine learning: A survey towards new research direction. *Scholars Journal of Engineering and Technology*, *1*, 56–65.

Dahal, K. R., Pokhrel, N. R., Gaire, S., Mahatara, S., Joshi, R. P., Gupta, A., Banjade, H. R., & Joshi, J. (2023). A comparative study on effect of news sentiment on stock price prediction with deep learning architecture. *PLOS ONE*, *18*(4), Article e0284695. https://doi.org/10.1371/journal.pone.0284695

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dong, Z., & Zhou, Y. (2024). A novel hybrid model for financial forecasting based on CEEMDAN-SE and ARIMA-CNN-LSTM. *Mathematics*, *12*(16), Article 2434. https://doi.org/10.3390/math12162434

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654–669. https://doi.org/10.1016/j.ejor.2017.11.054

Gao, P., Zhang, R., & Yang, X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, *25*(3), Article 53. https://doi.org/10.3390/mca25030053

Ghosh, B. P., Bhuiyan, M. S., Das, D., Nguyen, T. N., Jewel, R. M., Mia, M. T., Cao, D. M., & Shahid, R. (2024). Deep learning in stock market forecasting: Comparative analysis of neural network architectures across NSE and NYSE. *Journal of Computer Science and Technology Studies*, *6*(1), 68–75.

Ghosh, P., Neufeld, A., & Sahoo, J. K. (2022). Forecasting directional movements of stock prices for intraday

trading using LSTM and random forests. *Finance Research Letters*, *46*, Article 102280. https://doi.org/10.1016/j.frl.2021.102280

Guo, K., & Xie, H. (2024). Deep learning in finance assessing Twitter sentiment impact and prediction on stocks. *PeerJ Computer Science*, *10*, Article e2018. https://doi.org/10.7717/peerj-cs.2018

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Islam, M. S., Kabir, M. N., Ghani, N. A., Zamli, K. Z., Zulkifli, N. S. A., Rahman, M. M., & Moni, M. A. (2024). Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artificial Intelligence Review*, *57*(3), Article 62.

Ji, Y., Liew, A. W.-C., & Yang, L. (2021). A novel improved particle swarm optimization with long-short term memory hybrid model for stock indices forecast. *IEEE Access*, *9*, 23660–23671.

Jing, N., Wu, Z., & Wang, H. (2021). A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, *178*, Article 115019. https://doi.org/10.1016/j.eswa.2021.115019

Li, J.-C., Sun, L.-P., Wu, X., & Tao, C. (2025). Enhancing financial time series forecasting with hybrid deep learning: CEEMDAN-Informer-LSTM model. *Applied Soft Computing*, *177*, Article 113241. https://doi.org/10.1016/j.asoc.2025.113241

Li, Y., Chen, L., Sun, C., Liu, G., Chen, C., & Zhang, Y. (2024). Accurate stock price forecasting based on deep learning and hierarchical frequency decomposition. *IEEE Access*, *12*, 49878–49894.

Liapis, C. M., Karanikola, A., & Kotsiantis, S. (2023). Investigating deep stock market forecasting with sentiment analysis. *Entropy*, *25*(2), Article 219. https://doi.org/10.3390/e25020219

Lin, Y.-L., Lai, C.-J., & Pai, P.-F. (2022). Using deep learning techniques in forecasting stock markets by hybrid data with multilingual sentiment analysis. *Electronics*, *11*(21), Article 3513. https://doi.org/10.3390/electronics11213513

Liu, X., Guo, J., Wang, H., & Zhang, F. (2022). Prediction of stock market index based on ISSA-BP neural network. *Expert Systems with Applications*, *204*, Article 117604. https://doi.org/10.1016/j.eswa.2022.117604

Lu, W., Li, J., Li, Y., Sun, A., & Wang, J. (2020). A CNN-LSTM-based model to forecast stock prices. *Complexity*, *2020*, Article 6622927. https://doi.org/10.1155/2020/6622927

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University – Computer and Information Sciences*, *36*(4), Article 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Maqsood, H., Mehmood, I., Maqsood, M., Yasir, M., Afzal, S., Aadil, F., Selim, M. M., & Muhammad, K. (2020). A local and global event sentiment based efficient stock exchange forecasting using deep learning. *International Journal of Information Management*, *50*, 432–451. https://doi.org/10.1016/j.ijinfomgt.2019.07.011

Minh, D. L., Sadeghi-Niaraki, A., Huy, H. D., Min, K., & Moon, H. (2018). Deep learning approach for short-term stock trends prediction based on two-stream gated recurrent unit network. *IEEE Access*, *6*, 55392–55404. https://doi.org/10.1109/ACCESS.2018.2868970

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). Stock price prediction using news sentiment analysis. In *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications* (pp. 205–208). IEEE. https://doi.org/10.1109/BigDataService.2019.00035

Munyao, J. N., Oluoch, L. A., Iftikhar, H., & Rodrigues, P. C. (2025). Recurrent neural networks for hierarchical time series forecasting: An application to the S&P 500 market value. *Physica A: Statistical Mechanics and Its Applications*, *678*, Article 130869. https://doi.org/10.1016/j.physa.2025.130869

Mutinda, J. K., & Geletu, A. (2025). Stock market index prediction using CEEMDAN-LSTM-BPNN-decomposition ensemble model. *Journal of Applied Mathematics*, *2025*(1), Article 7706431.

Mutinda, J. K., & Langat, A. K. (2024). Stock price prediction using combined GARCH-AI models. *Scientific African*, *26*, Article e02374. https://doi.org/10.1016/j.sciaf.2024.e02374

Mutinda, J. K., Langat, A. K., & Mwalili, S. M. (2025). Forecasting Airtel stock prices through decomposition and integration: A novel VMD-GARCH-LSTM framework. *International Journal of Mathematics and Mathematical Sciences*, *2025*(1), Article 2710277.

UNCORRECTED PROOF-JHSN

Mutinda, J. K., & Yong, L. (2025a). Decomposition-ensemble approach for realized volatility prediction. *Computational Economics*. Advance online publication.

Mutinda, J. K., & Yong, L. (2025b). Deep learning applications in hierarchical time series forecasting of market value. *Computational Economics*. Advance online publication.

Mutinda, J. K., & Yong, L. (2026). Hybrid prediction framework using novel stability-enhanced dynamic thresholding feature selection and artificial intelligence methods for financial market trend prediction. *Computational Economics*. Advance online publication.

Niu, D., Sun, L., Yu, M., & Wang, K. (2022). Point and interval forecasting of ultra-short-term wind power based on a data-driven method and hybrid deep learning model. *Energy*, *254*, Article 124384. https://doi.org/10.1016/j.energy.2022.124384

Ouf, S., El Hawary, M., Aboutabl, A., & Adel, S. (2024). A deep learning-based LSTM for stock price prediction using Twitter sentiment analysis. *International Journal of Advanced Computer Science and Applications*, *15*(12).

Park, H. J., Kim, Y., & Kim, H. Y. (2022). Stock market forecasting using a multi-task approach integrating long short-term memory and the random forest framework. *Applied Soft Computing*, *114*, Article 108106. https://doi.org/10.1016/j.asoc.2021.108106

Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, *9*(24), Article 5574. https://doi.org/10.3390/app9245574

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications* (pp. 1394–1401). IEEE. https://doi.org/10.1109/ICMLA.2018.00227

Sivadasan, E. T., Mohana Sundaram, N., & Santhosh, R. (2024). Stock market forecasting using deep learning with long short-term memory and gated recurrent unit. *Soft Computing*, *28*(4), 3267–3282.

Usmani, S., & Shamsi, J. A. (2021). News sensitive stock market prediction: Literature review and suggestions. *PeerJ Computer Science*, *7*, Article e490. https://doi.org/10.7717/peerj-cs.490

Vuong, P. H., Phu, L. H., Van Nguyen, T. H., Duy, L. N., Bao, P. T., & Trinh, T. D. (2024). A bibliometric literature review of stock price forecasting: From statistical model to deep learning approach. *Science Progress*, *107*(1). https://doi.org/10.1177/00368504241236557

Yu, Y., Lin, Y., Hou, X., & Zhang, X. (2023). Novel optimization approach for realized volatility forecast of stock price index based on deep reinforcement learning model. *Expert Systems with Applications*, *233*, Article 120880. https://doi.org/10.1016/j.eswa.2023.120880

Zhang, G., Wang, J., Guo, H., & Zhang, X. (2018). The relationship between investor sentiment and stock market volatility: Based on the VAR model. In *WHICEB 2018 Proceedings* (Article 53). Association for Information Systems. https://aisel.aisnet.org/whiceb2018/53

UNCORRECTED PROOF-JHSN