International Journal of Mathematical, Engineering and Management Sciences
Vol. 11, No. 3, 1424-1443, 2026
https://doi.org/10.33889/IJMEMS.2026.11.3.058

Comparative Analysis of Time Series Forecasting Models for Predicting Hydrogen Fuel-Related Stock in the Indian Market

Angel Mary Jais
School of Chemical Engineering,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
E-mail: angelmary.jais2022@vitstudent.ac.in

Nandhini Haribabu Muthuvel
Department of Mathematics, School of Advanced Sciences,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
E-mail: nandhini.mh2022@vitstudent.ac.in

Sunanda Saha
Centre for Clean Environment,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
Corresponding author: sunanda.saha@vit.ac.in

Abhishek Das
Department of Mathematics, School of Advanced Sciences,
Vellore Institute of Technology, Vellore, Tamil Nadu, India.
E-mail: abhishek.das@vit.ac.in

Venkatesh Subramanian
Department of Biotechnology,
Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
E-mail: drvenkateshmsu@gmail.com

(Received on December 16, 2025; Revised on February 7, 2026 & April 12, 2026; Accepted on April 30, 2026)

Abstract
The study compares the capabilities of various time series and machine learning models including ARIMA, LSTM, CatBoost, XGBoost, and LightGBM, by predicting the equity movement for major Indian infrastructure and energy companies with hydrogen related exposure, namely Larsen & Toubro, NTPC Limited, JSW Energy Limited, and Adani Green Energy Limited. Hydrogen fuel is considered the most promising energy provider of the future, and an understanding of its position in the market is vital for its growth. The study uses historical data involving stock prices from April 2019 through April 2024 obtained from the National Stock Exchange of India. Using open price as the primary variable, the performance of the models is measured. Additional variables such as close price, highest price, lowest price, and volume are used for gradient boosting. Output graphs comparing actual prices and predicted prices are obtained. The results indicate that deep learning and gradient boosting outperform the statistical model. LSTM demonstrated the strongest short-term predictive accuracy through sequential learning among all models. Among the gradient boosting models, LightGBM provides consistent and robust performance by effectively capturing nonlinear feature interactions. Overall, the study highlights the growing importance of machine learning in interpreting India's renewable energy equity markets.

Keywords- Time series, LSTM, Gradient boosting, Stock market forecasting, Hydrogen fuel-related stock.

1. Introduction

The stock market is a transient financial system that plays an important role in the global economy. Prices may fluctuate rapidly and are governed by a variety of factors, including company performance, investor sentiment, geopolitical events, and macroeconomic conditions (Agarwal et al., 2014). Even slight price changes may significantly impact both individual investors and the overall market. These fluctuations have the potential to have a great impact, ranging from corporate decision-making to retirement portfolios. It is well established in the literature that forecasting the market prices may have some impact on better pattern understanding. However, stock market predictions have always proved challenging for experts due to the unpredictable nature of behaviour. It involves the use of various analytical techniques to forecast future price trends. While no model can offer perfect accuracy over long periods, these techniques offer valuable insights to investors and financial institutions. They help reduce risk, identify potential investment opportunities and improve decision-making. As one of the fastest growing economies in the world, India's energy requirements cannot be met by non-renewable energy alone; the limited supply of fossil fuels, the release of nitrogen oxides on the combustion of these fuels and the temperature surge due to the greenhouse effect collectively place non-renewable energy sources in a disadvantageous position (Sharma et al., 2021). This could strongly diminish the influence of non-renewable energy on the equity market in the near future. Therefore, countries are turning towards renewable energy as an appropriate replacement for the fossil fuels currently used by the majority. To shift towards a society with a lower carbon footprint, we can consider hydrogen fuel as the means to propel us towards this goal. This clean fuel, specially produced from renewable electricity and natural gas reforming, is a clean, highly efficient and high energy density alternative to fossil fuels. It has potential to be domestically produced fuel, increasing its desirability (Pareek et al., 2020). Hydrogen fuel exposed energy equities play an equally important role in the stock market. The transport and industry sectors are the key areas where these investments matter (Espegren et al., 2021). In 2009, an analysis of the Taiwan hydrogen economy was conducted, which led to the deduction that large investments were required to transfer a petroleum economy to a hydrogen one (Lee et al., 2009). Further, it was predicted that the real GDP growth rate for a hydrogen-based economy would be lower than that of a petroleum-based economy until 2020, due to the latter's established dominance; beyond 2021, with technological and infrastructural advancements, the hydrogen-based economy would outperform the other. This transition is slowly happening in India, with corporations investing in the growing hydrogen-based economy and allowing shareholders to benefit from this change. Over 80% of the hydrogen fuel produced is used in the automotive sector, while the remainder is used in the processing industry and energy sector. Export of hydrogen also plays a major role in economic growth. Companies from these territories will aim to grow their hydrogen fuel exposed stock. To get ahead of market competition, predictions can be employed by firms (Kovač et al., 2021). Previously, traditional models like ARIMA (AutoRegressive Integrated Moving Average) as well as deep learning models like LSTM (Long Short-Term Memory) were used to perform these predictions. However, GBM (Gradient Boosting Method) has recently taken the spotlight for such tasks. ARIMA is a traditional statistical model used for time series forecasting by analysing past values, trends, and patterns. It consists of three parts: AutoRegressive, which uses past values to predict the future; Integrated, which adjusts for trends and Moving Average, which reduces errors by considering past fluctuations. ARIMA works well for stable, linear trends (Mondal et al., 2014). LSTM is a deep learning model specifically designed for sequential data, like time series (Yadav et al., 2020). It shows the ability to remember long-term dependencies. This makes it particularly well-suited to model complex patterns in stock prices. XGBoost, CatBoost, and LightGBM are advanced machine learning models that are based on gradient boosting. Gradient boosting uses multiple decision trees sequentially to make predictions, improving them further at each iteration (Ma and Liu, 2008; Chandrika et al., 2023). There are many moving pieces to stock prices in the Indian hydrogen fuel sector, and the values can vary significantly based on regulatory policies, technology transitions and market sentiment (Shah et al., 2019;

Mehtab and Sen, 2020). GBM makes use of unseen data and patterns beyond the training set, and the ability to manage large feature spaces makes it useful for including numerous variables like company financials and volume trends. Thus, gradient boosting offers a powerful tool for predicting stock performance and identifying potential growth opportunities in the rising Indian hydrogen fuel market. These forecasting models are well suited to be applied on hydrogen-linked energy equities because such price series are shaped by a combination of persistent structural trends and abrupt regime shifts driven by market shocks. The strong relationship among financial variables shows the importance of the model which can learn complex, multivariate non-linear patterns such as the consideration of gradient-boosted tree ensembles. Since the daily prices are temporal dependence, sequential models like LSTM may provide an appropriate forecasting since they can capture nonlinear lag dynamics through gated memory. ARIMA is retained as a traditional baseline model to evaluate how well the linear differencing and autoregressive structure can explain the observed price movements compared to more flexible nonlinear models. However, GBM models are not widely used due to certain misconceptions; interpretability concerns and the highly complex nature of the market were always considered an obstacle for these types of models. With recent improvements in prediction model technology, they can be used in these scenarios. In this study, we focus on analysing the hydrogen fuel related equities of four prominent companies in the Indian clean energy sector: Larsen & Toubro, NTPC Limited, JSW Energy Limited, and Adani Green Energy Limited using various forecasting methods. The main contributions of this paper are:

(i) This research focuses specifically on hydrogen fuel-related energy stocks, an often-overlooked area of the market. The assumption that all forecasting models are well suited across all market stocks is required to be validated. Hence, through this study, we examine the effectiveness of a few models by specifically evaluating their performance on clean energy stocks in India.
(ii) The performance of the CatBoost, XGBoost, and LightGBM models is also analysed in this research. These models maintained efficient and streamlined model designs and also demonstrate their ability to deliver accurate results. The comprehensive comparison of prediction accuracy is enabled through the selection of these models from the wide range of available models.
(iii) Showcased whether GBM is a complex enough function to perform price prediction against traditional models. They were employed to introduce a volatility aware comparison of linear, sequential and tree ensemble forecasting frameworks on energy equity time series.
(iv) Specifically, we evaluate predictive performance using multiple error metrics (MAPE, RMSE, R^2) to distinguish short-term accuracy from robustness and point out the most practical model for this exact type of data.
(v) Provides methodological insight into sector-specific model transferability by demonstrating that forecasting performance hierarchies observed in general equity markets may shift when applied to hydrogen-focused energy equities, emphasising the need for sector-aware model selection strategies.

2. Research Dataset

For our study, the data[1] is retrieved from the National Stock Exchange of India Limited (NSE India) website, where specific company names are selected, and fetched the required data from historical data section. The website offers a platform that provides real-time and historical market data. In this research, historical data of every company for a duration of 1256 days starting from 11 April 2019 to 10 April 2024 are collected, with each date being assigned a serial number starting from 1 for representation purposes. All prices in the dataset are presented in Indian Rupees. The snippets of the datasets are presented in **Tables 1** to **4**, in which the opening price is the price at which the stock opened business for the period, the highest price is the highest value the stock attained, the lowest price is the lowest value the stock

[1] https://www.nseindia.com/

attained for the same period, and the closing price is the price of one share at the end of a given period. Volume is the sum of shares traded within a given time, both buys and sells.

Table 1. Larsen & Toubro stock prices.

Date	Open	High	Low	Close	Volume
2024 − 04 − 10	3800.00	3800.00	3750.10	3753.20	1,951,581
2024 − 04 − 09	3820.00	3827.95	3775.00	3785.25	1,997,057
2024 − 04 − 08	3745.00	3819.80	3743.10	3807.85	1,083,395
2024 − 04 − 05	3770.10	3790.00	3716.25	3743.10	1,645,597
2024 − 04 − 04	3804.90	3819.75	3731.05	3797.85	2,339,529

Table 2. NTPC Ltd stock prices.

Date	Open	High	Low	Close	Volume
2024 − 04 − 10	1920.05	1924.15	1895.00	1907.50	301,534
2024 − 04 − 09	1930.40	1945.00	1908.95	1916.20	294,088
2024 − 04 − 08	1948.05	1974.95	1904.40	1919.10	893,399
2024 − 04 − 05	1890.85	1937.75	1875.00	1901.95	724,439
2024 − 04 − 04	1900.50	1925.00	1882.00	1890.85	791,178

Table 3. Adani green energy limited stock prices.

Date	Open	High	Low	Close	Volume
2024 − 04 − 10	364.10	368.35	361.35	362.50	12,886,638
2024 − 04 − 09	365.00	366.90	360.35	362.90	10,010,148
2024 − 04 − 08	355.60	366.00	353.25	363.50	11,447,140
2024 − 04 − 05	356.40	358.20	351.55	354.55	11,208,130
2024 − 04 − 04	355.00	362.70	352.65	354.65	28,962,927

Table 4. JSW Ltd stock prices.

Date	Open	High	Low	Close	Volume
2024 − 04 − 10	615.50	630.00	612.05	616.05	3,433,168
2024 − 04 − 09	630.00	630.65	596.35	613.10	8,876,328
2024 − 04 − 08	602.00	632.65	601.95	627.50	6,476,004
2024 − 04 − 05	583.80	582.50	582.50	599.10	3,623,564
2024 − 04 − 04	582.00	577.10	577.10	583.05	8,023,364

It is important to note that the firms selected to collect this data are diversified energy and infrastructure companies rather than pure hydrogen producers. These stocks are treated as hydrogen linked equities of the company's energy divisions. This reflects investor sentiment and market expectations regarding hydrogen fuel adoption and clean-energy transition. Therefore, equity prices are used as an indirect proxy for hydrogen market dynamics, which is influenced by broader macroeconomic and company-specific factors. This limitation is acknowledged when interpreting forecasting performance and implications.

The key component considered for developing the model and performing the comparison is the opening prices of hydrogen fuel-related energy stocks, as it is often regarded as a market indicator that reflects the sudden reaction of the investors to overnight news, economic data releases, and global market movements. The opening price of the stocks is also unbiased and more consistent. Hence, considering the same variable across all forecasting models ensures a fair comparison, as it focuses on the predictive capabilities of each model. The ARIMA model depends on the univariate lagged relationships, and the LSTM network depends on temporal dependencies. It is to be noted that GBM models are additionally

supplied with auxiliary variables such as high, low, close, and volume, since tree-based ensemble models can effectively leverage multivariate input features to capture non-linear dependencies and feature interactions. This explicit multivariate information improves the prediction accuracy of the GBM model.

The Kernel Density Estimation (KDE) plots presented in **Figure 1** show the historical distribution of daily opening prices. The Adani Green Energy plot is highly multimodal with three main peaks, indicating that the stock's opening price has spent significant time periods clustering around several different valuation points. In contrast, the Larsen & Toubro distribution is predominantly unimodal and right-skewed, with a more stable core. The JSW Energy plot is distinctly pointing to a major structure shift where the stock moved from a low-price era to a higher one, spending considerable time at both. Finally, the NTPC Limited distribution is also multimodal, featuring significant groups at lower prices. The KDE slightly extends below zero in two plots due to its smoothing nature. Since stock prices cannot be negative, this portion does not represent actual data but results from the estimation bandwidth.

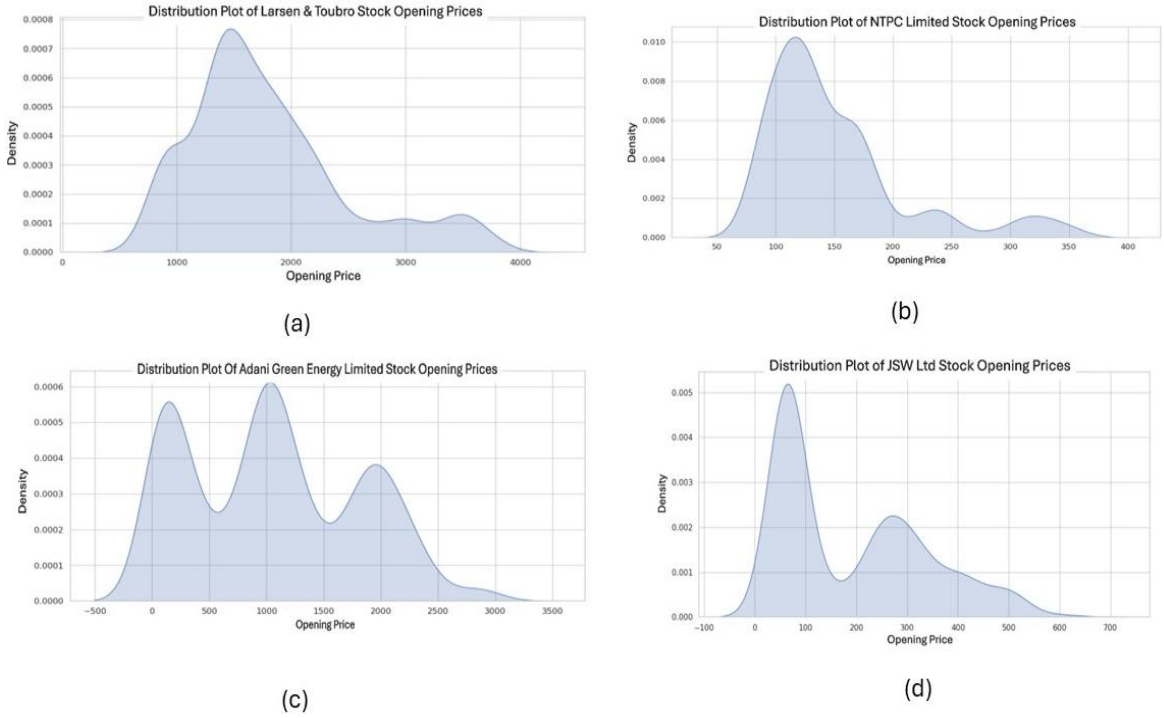


Figure 1. Distributions of open prices for Larsen & Toubro, NTPC limited, Adani green energy, and JSW Ltd stocks.

We start our inspection of correlations of the dataset variables to arrive at the study of the heatmaps (**Figure 2**). The correlation heatmaps detail the linear relationships among the five daily stock variables across the four listed companies. The most striking observation, seen across all four datasets, is the near-perfect positive correlation observed among the four price related variables. The coefficients show a very high level of redundancy. For example, the opening and closing prices of a stock on the same day are almost perfectly correlated, indicating that they move in complete synchronization. This shows a fundamental characteristic of daily stock data where the price range within a day is minimal relative to the overall price level. In contrast, the correlations involving the volume variable are significantly weaker and

heterogeneous across the companies. For Adani Green Energy and NTPC, the volume variable shows nearly zero correlations with stock prices. This indicates that the daily amount of trading activity is largely uncorrelated with the magnitude of price movements. Larsen & Toubro and JSW energy show a moderately negative correlation between volume and stock price variables. This suggests a tendency for higher stock prices to occur on days with lower trading volumes. This collective independence of volume from the price variables plays a vital role, since it identifies the sole non-redundant feature in the dataset. This contributes to the unique market activity data that is not already captured by the highly correlated price variables, making it more essential for GBM models.

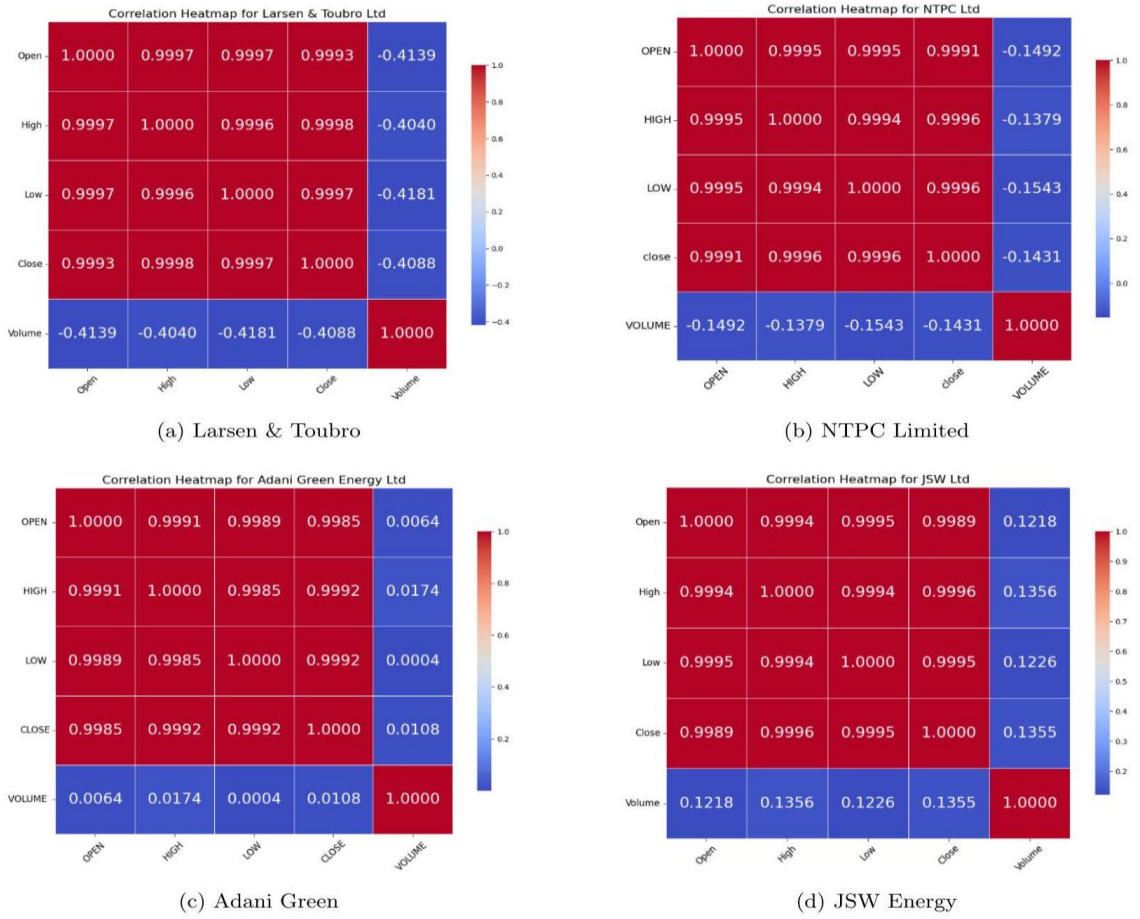
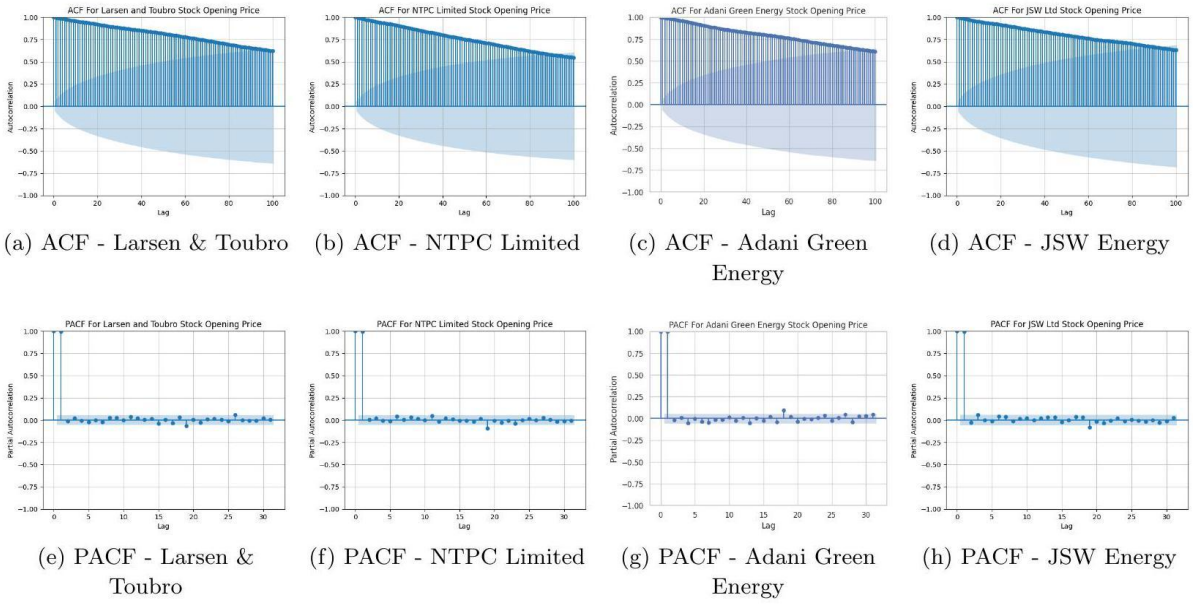


Figure 2. Correlation heatmaps of Larsen & Toubro, NTPC limited, Adani green, and JSW energy.

Autocorrelation function (ACF) essentially measures the association between the stock's opening price today and the opening price of the stock on prior days (**Figure 3**). The autocorrelation coefficient ranges from -1 to 1; the closer it is to one, the stronger the positive correlation and the closer it is to negative one, the stronger the negative relationship. Partial autocorrelation function (PACF) refines this concept. Instead of simply checking if today's price mirrors previous days, PACF specifically examines the direct relationship between today's price and earlier values, while removing the influence of any intermediate days. This gives a better understanding of how many lags prior to the requisite correlations are in the data. Looking at the ACF plots for these stock datasets (**Figure 3**), we see that the autocorrelation begins at 1

for lag 0, as it rightly should, because a variable should always be perfectly correlated with itself. The autocorrelation values have a downwards trend as the lag increases. The correlating current and lagged opening prices decline. This decrease in correlation suggests that the two most recent prices have greater predictive capabilities for the near future. However, this trend will lose its predictive capability quickly as we see further lags. The PACF plots for Larsen & Toubro, NTPC Ltd, JSW Energy Ltd, and Adani Green Energy Ltd lend more transparency to these relationships. Each bar on the PACF plot represents the partial autocorrelation at a specific lag, with a shaded confidence interval around zero. Values extending beyond this region (notably at lags 1 and 2) are statistically significant, suggesting that the opening price is still meaningfully connected to those from two days prior. Beyond lag 2, partial autocorrelations fall within the confidence interval, implying no significant direct correlation with opening prices more than two days ago.



(a) ACF - Larsen & Toubro (b) ACF - NTPC Limited (c) ACF - Adani Green Energy (d) ACF - JSW Energy

(e) PACF - Larsen & Toubro (f) PACF - NTPC Limited (g) PACF - Adani Green Energy (h) PACF - JSW Energy

Figure 3. ACF and PACF plots of the stock open prices: Larsen & Toubro, NTPC limited, Adani green energy, and JSW energy Ltd. The first row displays the ACF plots, while the second row shows the corresponding PACF plots.

3. Methodology and Implementation

Models

The primary objective of this section is to present a comprehensive understanding of the five models used in this study: ARIMA, LSTM, XGBoost, LightGBM, and CatBoost. Specifically, the application of ARIMA and LSTM for forecasting hydrogen fuel stock prices in India remains unexplored. Each model employs different methods to address stock price prediction challenges. The typical first step in a time series analysis is to determine if the data is stationary. This generally starts with an examination of the raw data to determine if there are any confirmed trends or seasonal components that can be eliminated via differencing. The model can be differentiated based on the degree of non-stationarity; in some cases, any non-stationary data is best modelled via first-order differencing, or a second-order difference may yield better results. In addition to differencing to eliminate seasonality or trend, the data may also need to be smoothed and reduced for noise with a moving average or filtering technique to uncover the underlying

pattern present in the time series. Where relevant, other variables (e.g., local market conditions, environmental factors) can be included as exogenous variables, much like a weather forecast.

3.1 ARIMA

The ARIMA model has three components, which are the order of autoregression (p), order of differencing (d) and order of moving average (q). The autoregressive component incorporates a maximum number of past values up to p lags. The moving average part considers dependencies on lagged forecast errors up to q steps, and the integrated part makes sure the time series is stationary after differencing d times. The ARIMA process iteratively combines these processes into a forecast based on both past observations and future estimates. The model for ARIMA (p, d, q) is defined as (Sirisha et al., 2022):

$$\Phi(B)(1-B)^d X_t = \Theta(B)\epsilon_t \quad (1)$$

where,

- $X_t =$ actual time series value at time t
- $\epsilon_t =$ random error term at time t
- $B =$ backshift (lag) operator, $BX_t = X_{t-1}$
- $\Phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i =$ autoregressive operator polynomial
- $\Theta(B) = 1 + \sum_{j=1}^{q} \Theta_j B^j =$ moving average operator polynomial

This type of analysis permits adaptive modelling of time-based sequential changes in the data while retaining historical influence from the past, thereby avoiding unrealistic expectations about future behaviour. The typical loss function in this case is the Mean Squared Error (MSE), which indicates how well the forecast is performing. implemented with a fixed order of (0,1,0) as a baseline differencing model for non-stationary series. The optimization algorithms will try the parameter settings until they converge, or they may stop early if no improvement is evident. However, ARIMA assumes linear dependence and requires stationarity after differencing, which limits its ability to capture abrupt nonlinear market shifts and volatility clustering commonly seen in equity pricing.

3.2 LSTM

As only priming data for our LSTM, the first preprocessing step is to chop up the time series data into sliding sequences. Every sequence will generally include a constant number of previous time steps in order to predict the probable target value. Normalization will be used to scale all features between 0 and 1, which makes training more stable and efficient and helps with vanishing/exploding gradient issues. In addition to time series features, additional data characteristics could be added to serve as additional context that could help model performance (Jarrah and Derbali, 2023; Sun and Tian, 2023). The LSTM model architecture is specifically developed for handling long-term dependencies in sequential data through a unique cell structure with three main gates: the input gate, which decides what new information to accept; the forget gate, which decides what information to forget; and the output gate, which decides what information to send to the next cells in the network (Rezaei et al., 2021). The number of units can be interpreted as the LSTM's capacity to learn complex patterns. A typical LSTM network will have stacked LSTM layers that are structured to retain temporal context, followed by a fully connected and dense layer, for which activation function choices include ReLU or linear, depending on the prediction task. The last output of the model is the value to be predicted. The LSTM model, like all neural networks, contains settings and states explained using specific mathematical equations, which describe transformations and gating as detailed below:

Forget gate:

$$\mathfrak{f}_t = \sigma_g(\mathcal{W}_f \cdot [\mathfrak{h}_{t-1}, x_t] + \mathfrak{b}_f) \quad (2)$$

Memory gate:

$$\mathfrak{i}_t = \sigma_g(\mathcal{W}_i \cdot [\mathfrak{h}_{t-1}, x_t] + \mathfrak{b}_i) \quad (3)$$

Temporary cell gate:

$$\tilde{\mathfrak{C}}_t = \tanh(\mathcal{W}_c \cdot [\mathfrak{h}_{t-1}, x_t] + \mathfrak{b}_c) \quad (4)$$

Current cell state:

$$\mathfrak{C}_t = \mathfrak{f}_t * \mathfrak{C}_{t-1} + \mathfrak{i}_t * \tilde{\mathfrak{C}}_t \quad (5)$$

Output gate:

$$\mathfrak{O}_t = \sigma_g(\mathcal{W}_o \cdot [\mathfrak{h}_{t-1}, x_t] + \mathfrak{b}_o) \quad (6)$$

Hidden state:

$$\mathfrak{h}_t = \mathfrak{O}_t * \tanh(\mathfrak{C}_t) \quad (7)$$

where,

- x_t: Input vector at time t
- $\mathcal{W}_f, \mathcal{W}_i, \mathcal{W}_c, \mathcal{W}_o$: Weight matrices
- $\mathfrak{b}_f, \mathfrak{b}_i, \mathfrak{b}_c, \mathfrak{b}_o$: Bias terms associated with each respective gate
- $\mathfrak{f}_t$: Forget gate activation
- $\mathfrak{i}_t$: Input (memory) gate activation
- $\tilde{\mathfrak{C}}_t$: Temporary cell gate output
- $\mathfrak{C}_t$: Cell state
- $\mathfrak{O}_t$: Output gate activation
- $\mathfrak{h}_t$: Hidden state
- $\sigma_g(\cdot)$: Sigmoid activation function
- *: Element-wise multiplication operator

A candidate cell state produces potential updates to the cell's current memory. The cell state combines old and new data to handle long-term connections. This set-up also allows important patterns to stick around. The output gate then determines what part of the cell state is sent to the 'next' hidden state, which is passed along to the next LSTM unit in the sequence. x is the current input at each timestep. The weight matrices and the bias terms are all required to parametrize the operation in each gate. The activation functions are used to convert the output into a probability like value which involves gating mechanisms. The interaction of these components allows LSTM to maintain important information over long sequences and remove residues. Both distant and recent dependencies are significant in time series, so this ability is advantageous. The Adam optimizer is commonly used because it implements an adaptive learning rate. It is implemented in this study. Fairly typical parameters could include a learning rate of 0.001 and a batch size of 32. Training takes place over a set number of epochs, and early stopping is implemented if the validation loss does not improve in the set number of epochs. If overfitting occurs, training should halt at that point. Although an LSTM model is well-suited for tracking non-linear sequential dependencies, it has its drawbacks, including higher computational costs and overfitting with limited data. They generalize noisy data to an extreme.

3.3 CatBoost

When using CatBoost, the preparation of the data begins with importing and formatting all categorical and numerical features needed for the model. CatBoost handles categorical features, so it is not required to do much preparatory work (Ben Jabeur et al., 2021). The model split the data into training, validation, and testing sets. CatBoost has its own way of encoding categorical variables using target stats, which makes the process easier and cuts down on target leakage risks. It also handles missing values; however, the scaling or normalizing of the numeric features may still be required in some scenarios. Other than that, the CatBoost algorithm is robust enough to work well even without extensive scaling. The objective function is defined as:

$$Obj = \sum_{i=1}^{n} \beta(y_i, \hat{y}_i) + \Gamma(f_k) \quad (8)$$

where,

- y_i: True target value of the i^{th} sample
- $\hat{y}_i$: Predicted value of the i^{th} sample
- $\beta(y_i, \hat{y}_i)$: Differentiable loss function between actual and predicted values
- $\Gamma(.)$: Regularization term
- f_k: Individual tree in the ensemble model
- n: Number of training samples

CatBoost's main innovation is its 'ordered boosting', which is a way to solve the overfitting issues that come with traditional gradient boosting while simultaneously improving its generalization from better permutation handling. CatBoost builds symmetric decision trees, so when it grows trees, it grows all the leaves, settling at one depth. This design choice makes the training process stable and efficient. In CatBoost, model training usually occurs using loss functions such as MSE when modelling in a regression context. The training of the model continues by optimizing using gradient descent. The power of CatBoost is further pushed by the option of GPU, which can halve the training time of a CPU-only approach. However, the model complexity may result in longer training times than much simpler algorithms, and it is also unable to explain causality.

3.4 XGBoost

Preprocessing is an essential first step when using XGBoost. Categorical features need to be converted to numeric format by using one-hot encoding or label encoding. Scaling or normalizing the dataset is a good idea as well to have similar ranges for features to encourage optimal algorithm performance (Su et al., 2022). While XGBoost can internally handle missing values, filling in the gaps in advance will generally help improve training time. Feature engineering can contribute significantly to improving model performance (Chang et al., 2018). XGBoost itself is a highly optimized implementation of gradient-boosted decision trees. Unlike a standard decision tree, we sequentially build the trees, and each successive tree attempts to rectify the mistakes of the previous trees (Yuan et al., 2021; Dong et al., 2023). Another important aspect of XGBoost is the use of L1 and L2 regularization to prevent overfitting-a common problem in machine learning. XGBoost also uses a second-order approximation for the loss function-this increases training speed and accuracy from regular boosting implementations. XGBoost can produce optimistic results if temporal leakage occurs through random splitting, and it requires careful validation design for time-dependent data. The objective function is defined as:

$$Obj = \sum_{i=1}^{n} \beta(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Gamma(f_k) \quad (9)$$

$$\Gamma(f_k) = \gamma T_k + \frac{1}{2}\lambda \sum_{j=1}^{T_k} \omega_{kj}^2 \quad (10)$$

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \quad (11)$$

where,

- f_k: k^{th} regression tree (base learner)
- $\mathcal{F}$: Functional space/set of all possible trees
- K: Total number of boosting trees in the model
- T_k: Number of leaves in tree k
- ω_j: Leaf weight of j^{th} leaf node
- γ: Regularization parameter penalizing number of leaves
- λ: Regularization parameter for L2 penalty on leaf weights
- x_i: Input features of sample i

The other parameters are the same as those defined in the previous subsections (Sections 3.1-3.3).

3.5 LightGBM

Similar to the other boosting algorithms, LightGBM requires the preprocessing of data before developing the model. This generally involves the changing of categorical variables into numbers through label encoding, addressing the missing values appropriately, and normalizing the numerical feature to ensure the consistency. During time series prediction, developing lag variables and trend features may improve the model's performance to capture latent structure (Hartanto et al., 2023). A unique feature of LightGBM is the way it grows trees in a leaf-wise manner. Instead of finding splits for nodes level by level, LightGBM explores each leaf and finds the leaf that will result in the greatest loss reduction at each round of boosting. This allows LightGBM to construct deeper trees in return, which is helpful for finding complex structures in data and thereby making very fine resolutions of predictions. However, increased depth of trees can also increase the risk of overfitting, particularly with smaller or noisier datasets. LightGBM also draws upon its hallmark efficiency. By using the histogram-based learning method and feature bundling, LightGBM is faster to train and also uses less memory than typical implementations of gradient boosting even on large datasets. Model optimization performance can continue by tuning hyperparameters (learning rate, maximum depth, boosting iterations, etc.). Regularization and early stopping are also extremely relevant to improve generalizability and reduce overfitting (Ke et al., 2017; Chen et al., 2019). While LightGBM is a strong combination of speed and accuracy, parameter tuning often is critical to yield reliable results across any data environment. Optimizing hyperparameter selection is very important SML models don't just perform on a hold-out data sample, hyperparameter tuning also optimizes runtime efficiency too. LightGBM was implemented using the default LGBMRegressor configuration. Temporal information was incorporated through engineered date-derived features. The model was evaluated using an 80-20 random hold-out split. LightGBM's leaf-wise tree growth can overfit if depth and leaf constraints are not controlled, especially under noisy or highly volatile market regimes. Performance may depend strongly on how time and lag information is engineered into the feature set. The objective function is defined as:

$$Obj^{(t)} = \sum_{i=1}^{n} \beta\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Gamma(f_t) \quad (12)$$

Expanding the function using a second-order Taylor approximation and solving for a tree with j-th leave, the optimal weight function for that specific leave can be determined. Finally, the optimal value of the objective function for that tree is:

$$Obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{J} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i+\lambda} + \gamma J \quad (13)$$

where,

- g_i: First derivative (gradient) of loss w.r.t. $\hat{y}_i^{(t-1)}$
- h_i: Second derivative (Hessian) of loss w.r.t. $\hat{y}_i^{(t-1)}$
- I_j: Set of data indices in leaf j
- γ: Penalty term for the number of leaves J
- J: Total number of leaves in the tree

The other parameters are the same as those defined in the previous subsections (Sections 3.1-3.4).

3.6 Hyperparameter Tuning

Hyperparameter tuning is a crucial step for all forecasting techniques. Hyperparameters were selected using a combination of default recommended values from literature and tuning to balance predictive performance and efficiency. Significant adjustments are required depending on the dataset and the problem to be solved. Different train-test splitting strategies were employed based on the mathematical structure of the models. ARIMA and LSTM are sequential models; a chronological 80-20 split was used. This approach is used to preserve temporal ordering and prevent information leakage, as these models explicitly rely on past observations to predict. A fixed-window validation strategy was adopted without using rolling or walk-forward retraining. ARIMA used a lag-based look back window. The look-back window length for LSTM was selected based on empirical experimentation across multiple window sizes (e.g., 5, 10, 20) with 10 providing the best balance and is a commonly used value to study financial data. In contrast, gradient boosting models (Zhou et al., 2019) (LightGBM, XGBoost, and CatBoost) are treated as tabular regression models (**Table 5**). For these approaches, an 80-20 random split was adopted. Temporal dependencies were incorporated through explicit feature engineering of date-related variables. While this split strategy aligns with the theoretical assumptions of each model class, the use of random splitting for tree-based models may lead to optimistic performance, which is a constraint. All models were evaluated at the same target variable (daily opening price). Although the input format differs by model class, performance metrics were computed on held-out observations. A one-step-ahead forecast horizon was used, where models were trained to predict the next trading day's opening price based on historical observations.

Table 5. Key hyperparameters used for each model.

Model	Key hyperparameters
ARIMA	(p, d, q) = (0, 1, 0)
LSTM	Look-back window = 10; 2 LSTM layers; 100 units each; batch size = 32; epochs = 100; optimizer = Adam
LightGBM	Learning rate = 0.1; estimators = 100; max depth = unlimited (-1); leaves = 31
XGBoost	Learning rate = 0.3; max depth = 6; estimators = 100
CatBoost	Iterations = 100; tree depth = 2; native categorical feature handling

4. Discussions

4.1 ARIMA Model

The ARIMA model's performance is evaluated by comparing predicted stock open prices against actual prices for the four companies (**Figure 4**). The analysis reveals a pattern. The model performs well in stable trending periods but struggles quite a lot during periods of high volatility and sharp changes in the

market. For NTPC Ltd and JSW Energy Ltd, the predictions of the ARIMA model are mostly align with the actual price line. It tracks the general long-term price trend with accuracy and reveals the model's moderate success when the market is on a smooth run. However, as market volatility increases, the constraints become more obvious. Both the graphs for Larsen & Toubro and Adani Green Energy Ltd fall short of predicting the actual price movements, especially during local peaks and severe declines. Notably, this is most pronounced in the case of Adani Green Energy (**Figure 4(c)**), which produces abrupt price movements too complex for the model to capture. This results in substantially greater short-term differences between the predicted and actual values. A similar pattern is observed for Larsen & Toubro, which had greater absolute errors compared to the NTPC chart but more moderate errors during periods of volatility. This shows that the ARIMA model is effective at capturing long-term trends and stable patterns in stock prices. However, since its basic structure relies on historical lags and linear relationships, it cannot adapt quickly to non-linear and abrupt market fluctuations. This results in a performance lag and reduces predictive accuracy. Hence, this model is suitable for relatively stable market conditions.

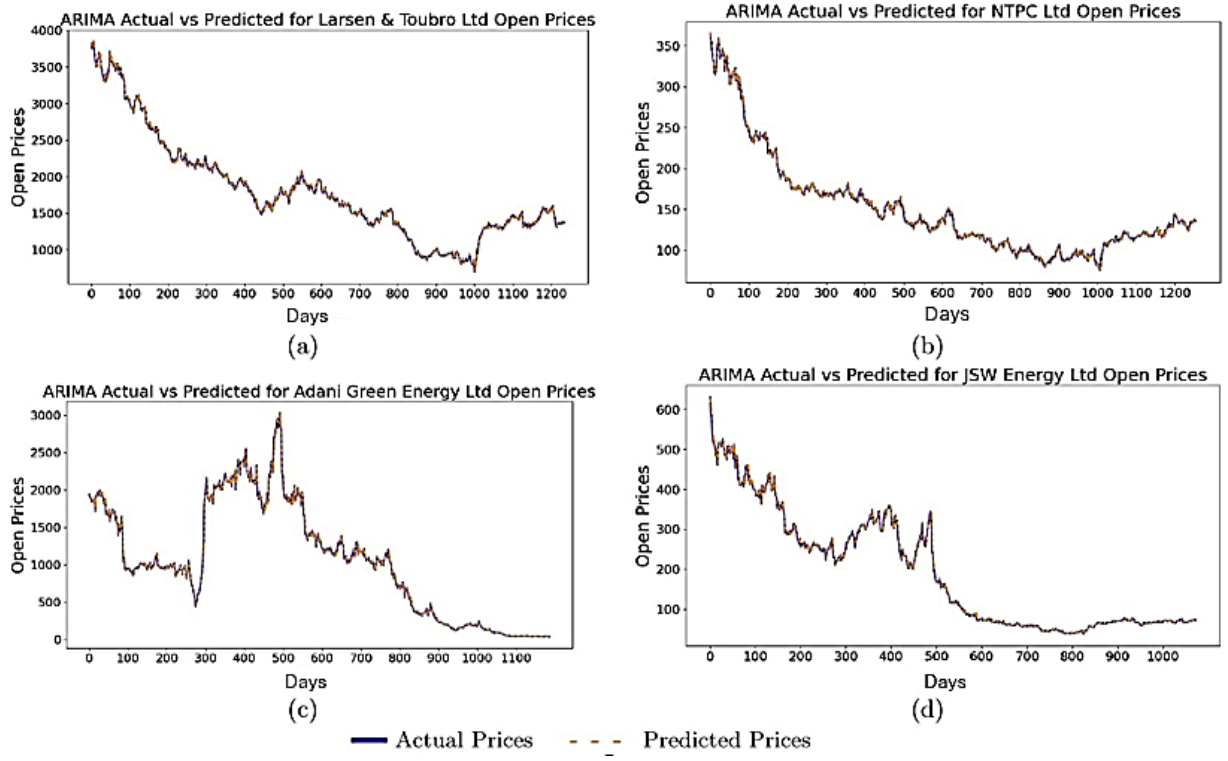


Figure 4. Actual vs predicted stock open prices for hydrogen fuel-related companies using ARIMA model.

4.2 LSTM Model

The performance of the LSTM model on the given four datasets shows that it has the capability to handle the non-linear trends properly; however, its sensitivity to the extreme volatility is clearly visible (**Figure 5**). For NTPC Ltd and JSW Energy Ltd, the LSTM model demonstrated excellent accuracy. The test and training lines are indistinguishable. This shows the model's understanding of the overall downward trend and the subtle short-term fluctuations within the training data. More importantly, the test line for both prices closely follows the prices at the end of the time series, indicating strong generalization outside the sample during periods of low-price change (Hochreiter and Schmidhuber, 1997; Gülmez, 2023). The

model maintains a generally strong performance for Larsen & Toubro, in which the predictions closely track the long-term price decline. While there is an expected deviation during short-term volatile swings where the model slightly lags, the line covers the overall market movement efficiently. On the other hand, Adani Green Energy Ltd presented the most complex scenario. Throughout the historical period of volatility, the LSTM predictions showed a prominent divergence from the actual prices, specifically during the estimation of the highest peaks and lowest troughs of the stock price. Hence, this model's inability to accurately locate the size and timing of the abrupt price changes resulted in a larger prediction lag, leading to underperformance of the model for this stock.

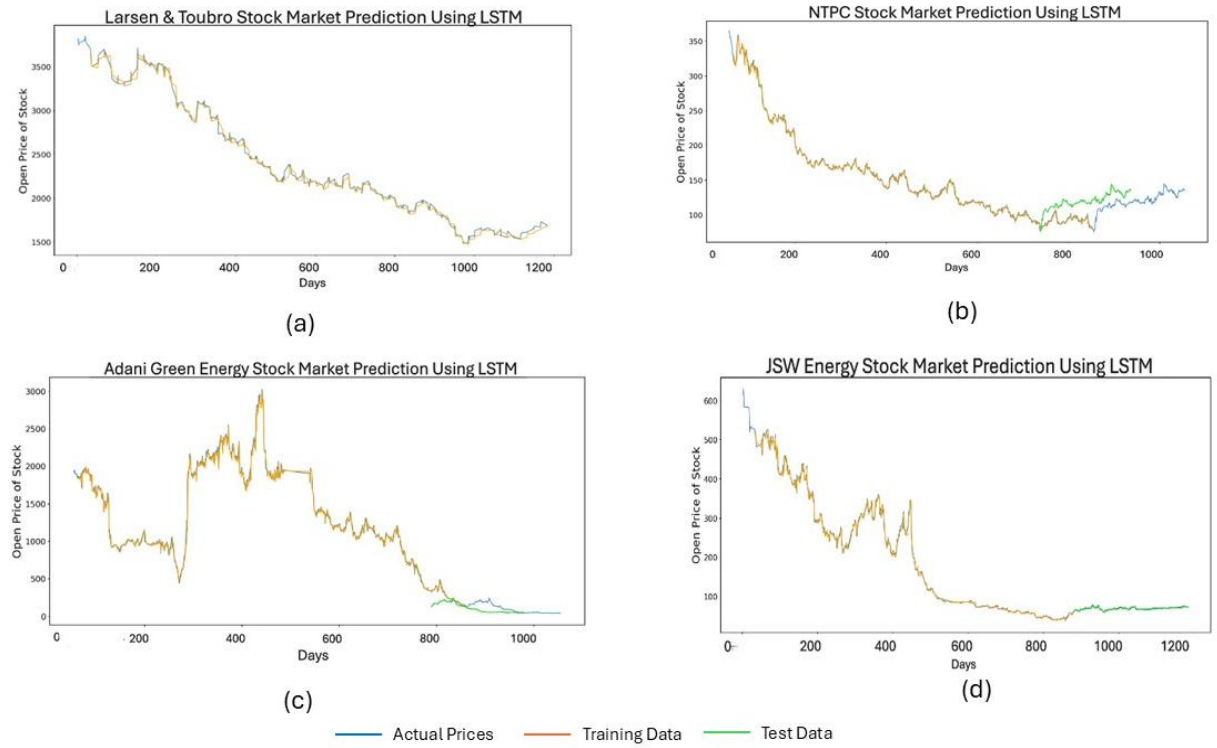


Figure 5. Stock market prediction using LSTM.

4.3 Gradient Boosting Models

When assessing the GBM model's effectiveness in forecasting daily prices, there is a noticeable drop in accuracy with increasing stock volatility (**Figure 6**). For NTPC Ltd. and JSW Energy Ltd., both of which have similar moderate volatility characteristics, the model achieves the highest accuracy. The forecasted values from CatBoost, XGBoost, and LightGBM are generally well aligned with price peaks and valleys across the sample index, showing that the models are effectively capturing short-term high-frequency price series patterns. Prediction errors are typically small, indicating that these models are stable and reliable at forecasting for stocks with less volatility. The predictive capability slightly diminishes for Larsen & Toubro and Adani Green Energy. While the models generally track the overall movement, the deviations become more visible during periods of sharp upward and downward spikes. This indicates that the model struggles more to forecast the rapid, large-magnitude anomalies in these moderately volatile datasets. The ensemble models demonstrate their strength by identifying the general height of the peaks and depths of the troughs, the magnitude of the vertical distance between the value and the predicted

value is greater, especially during high-volatility events. This confirms that even advanced non-linear models are heavily affected by the extreme instability and sudden price changes in highly volatile assets (Shrivastav and Kumar, 2022). Consequently, their forecasts capture interactions among multiple correlated variables and do not reflect purely univariate patterns (Ahn et al., 2023). They introduced a Sample Index axis of short-term and daily prediction, framing each day's prediction as an independent regression problem and visually emphasizing the short-term, point-by-point prediction accuracy.

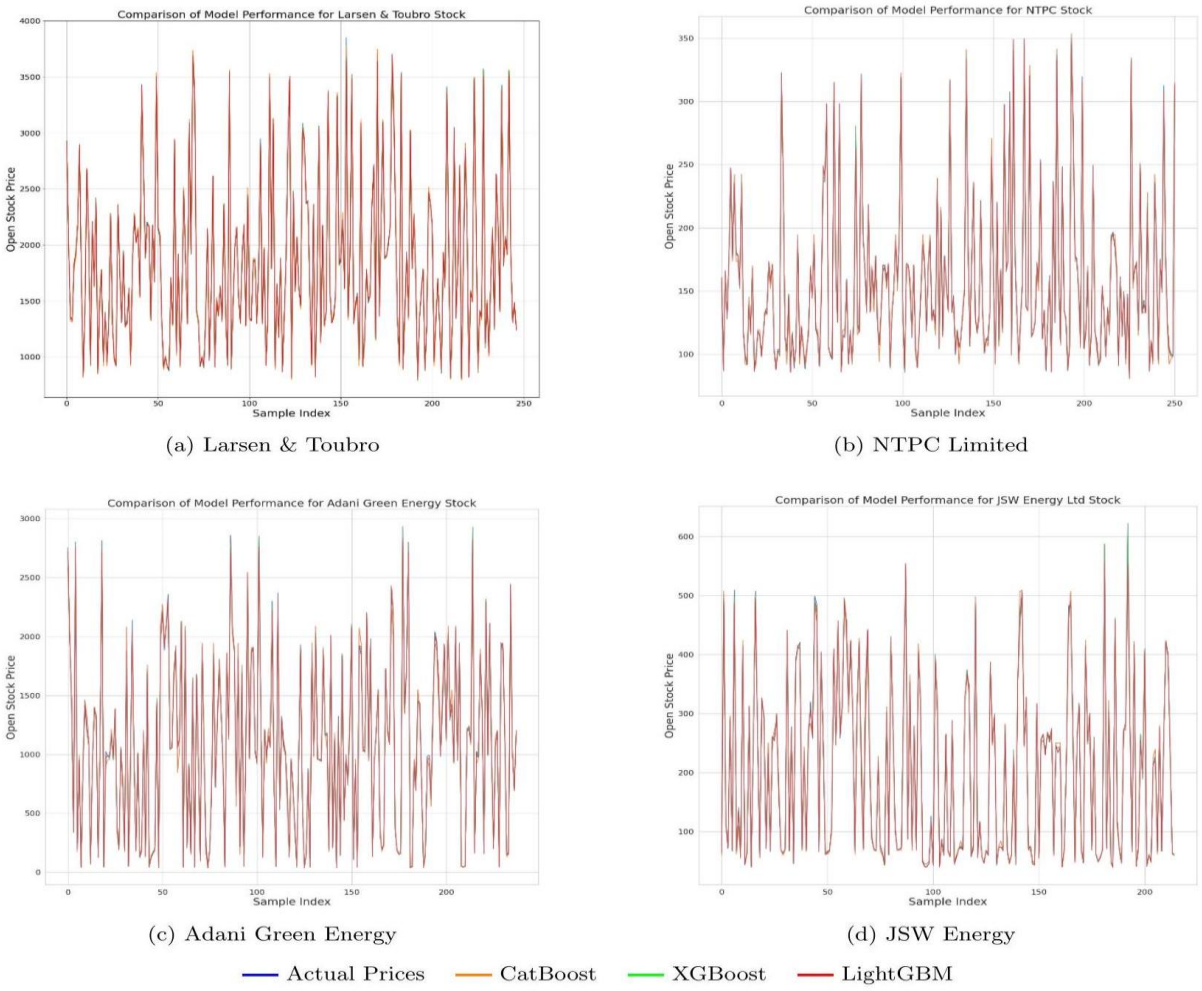


Figure 6. Comparison of model performance using gradient boosting.

Both the ARIMA and LSTM models were trained in the same historical "Open Price" data of all four companies. Minor visual differences in the actual price curves arise from preprocessing steps required by deep learning models, such as scaling and sequence windowing. The GBM models also exhibit variations in their plotted curves. This is because GBM models incorporate additional input features such as high, low, close prices and volume along with the open price. Despite the visual differences, the model environments and evaluation periods remain the same, and the cross-comparison fairness and consistency are guaranteed by all models being based on the same base and evaluation datasets, with the same performance comparison metrics.

Accuracy of forecasting

Evaluation of the performance of each model was conducted through the analysis of several error metrics. These were: Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and the Coefficient of Determination (R^2) for each model. MAPE shows the average percentage deviation between observed and predicted values. A smaller MAPE indicates stronger model performance in percentage terms. This is because it reflects that the predicted values are closer to the actual values, as expressed in the following equation:

$$\text{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (14)$$

RMSE is calculated by taking the differences between predicted and actual values, squaring those differences, averaging them, and then taking the square root of that average. This metric is sensitive to outliers because of the squaring step. In general, a lower RMSE suggests that the model's predictions are closer to the actual values, indicating better predictive performance equation:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (15)$$

$$R^2 = 1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (16)$$

The coefficient of determination (R^2) given by Equation (16) measures the extent to which variability in the dependent variable can be explained by the independent variables within a regression model. When R^2 is close to 1, it suggests that the model accounts for most of the variance observed in the data. This metric is fundamental for assessing the explanatory power of a regression analysis.

***Notations*:**

- y_i - Observed value for the i^{th} data point
- $\hat{y}_i$ - Predicted value for the i^{th} data point
- $\bar{y}$ - Mean of the observed values
- n - Total number of data points

5. Results

Table 6. MAPE values for different models.

Dataset/Model	CatBoost	XGBoost	LightGBM	ARIMA	LSTM
Larsen & Toubro	1.29	0.82	0.86	1.36	0.018
NTPC Ltd	2.68	0.97	0.99	1.44	0.015
Adani Green Energy	6.33	3.22	1.79	2.39	0.18
JSW Ltd	1.64	1.59	1.35	2.37	1.13
Average	2.98	1.65	1.24	1.89	0.33

Table 7. RMSE values for different models.

Dataset/Model	CatBoost	XGBoost	LightGBM	ARIMA	LSTM
Larsen & Toubro	28.94	23.96	24.55	10.68	32.13
NTPC Ltd	62.52	36.61	38.45	17.22	2.50
Adani Green Energy	5.19	2.57	2.32	74.39	1.29
JSW Ltd	2.07	5.65	6.04	23.30	1.93
Average	24.68	17.19	17.84	31.39	9.46

Table 8. R^2 values for different models.

Dataset/Model	CatBoost	XGBoost	LightGBM	ARIMA	LSTM
Larsen & Toubro	0.9985	0.9989	0.9994	0.9744	0.9676
NTPC Ltd	0.9993	0.9996	0.9994	0.9541	0.9531
Adani Green Energy	0.9997	0.9996	0.9996	0.9900	0.9550
JSW Ltd	0.9947	0.9985	0.9991	0.9802	0.7646
Average	0.9980	0.9991	0.9993	0.9746	0.9100

The comparative evaluation of ARIMA, LSTM, and Gradient Boosting models demonstrates the clear differences in predictive error behaviour across hydrogen fuel related energy companies. Based on average error metrics, LSTM consistently achieved the lowest MAPE and RMSE values across datasets (**Tables 6** and **7**). This shows the ability of sequential deep learning architectures to effectively capture short-term temporal dependencies and price movement patterns present in daily equity time series. The high R^2 values observed are due to the strong autocorrelation and continuity present among the prices (**Table 8**). For tree-based models, the random hold out splits may have yielded overtly optimistic results. ARIMA and LSTM are more trend following types of models which can lead to overfitting behaviour. Among the Gradient Boosting models, LightGBM and XGBoost both demonstrated strong predictive performance. LightGBM emerged as the best performing model within the GBM category, showing slightly stronger variance explanation compared to XGBoost. R^2 values were only used as a secondary differentiator between the boosting models. LightGBM demonstrated a more balanced performance profile, combining strong predictive accuracy with greater robustness across varying volatility regimes, making it a reliable GBM choice for real-world financial forecasting deployment. The ARIMA model is the least desirable model among the five. It lagged during periods of rapid fluctuation. Its relatively higher RMSE values show that the model finds difficulty in adapting to sharp market movement. This is a known limitation of linear statistical forecasting in a non-linear, highly volatile environment. CatBoost performed poorly across the metrics for all datasets. Higher prediction errors observed in companies such as Adani Green Energy and JSW Energy Ltd can be attributed to increased price volatility and irregular trading behaviour. This introduces higher short-term noise and reduces the model's predicting ability. These findings confirm the fundamental challenge that volatility introduces to financial forecasting, irrespective of the model's complexity. LSTM models tend to be the ones that perform better in smoother price series, where continuity and short-term sequential dependencies are stable. This allows the network to effectively learn localized time patterns. While GBM models remain robust across varying volatility conditions, their accuracy is considerably lower. In general stock market analysis, machine learning is the preferred model in contrast to these results (Nakagawa and Yoshida, 2022). The lower average prediction errors observed for LSTM in this study indicate that sequential learning provides an advantage over tree-based machine learning models for hydrogen-aligned energy stock forecasting. Overall, LSTM emerges as the best performing model in terms of raw predictive accuracy, making it the most accurate model for forecasting stock price movements in this study.

6. Conclusion

This study conducted a comparative analysis of traditional statistical and advanced machine learning models to forecast the daily opening prices of four major Indian energy companies. The influence of the volatile nature of equity and its generally inverse relation to accuracy is observed throughout the study. Highly unstable assets produced errors across all approaches. To reconcile comparative findings, performance is measured by two complementary criteria, short-term predictive accuracy and robustness. LSTM demonstrates strong short-term accuracy shown by lower RMSE and MAPE values. Its ability to learn sequential dependencies from time-ordered inputs plays a vital role in achieving high accuracy. On an added note, on GBM, LightGBM provides stable performance across all the datasets by leveraging

multivariate feature interactions. The results also provide insight into model transferability across financial sectors. The observed variation in model performance across hydrogen-focused energy equities suggests that forecasting models developed and validated on general market datasets may not consistently retain their performance hierarchy when applied to niche, policy-driven sectors. This reinforces the importance of sector-specific validation before deploying forecasting models in emerging clean energy markets, particularly where multiple factors strongly influence price behaviour.

This information can be used by firms, stakeholders, and even ordinary investors to increase equity output to a limit. In the rapidly evolving market conditions, these two models are well-suited for forecasting applications in the Indian renewable energy sector. The credibility of gradient boosting is also established. Further usage of LightGBM can be encouraged. The findings also highlight that extreme market volatility is still a challenge for all forecasting frameworks. The hybrid architectures such as Bi-LSTM, GRU, and deep ensemble models need to be explored for dynamic financial environments. This study offers a practical reference that may guide future researchers to apply these methods effectively in financial forecasting and related data-driven areas.

Conflicts of Interest

The authors confirm that there are no conflicts of interest associated with this work.

Acknowledgments

The authors acknowledge the financial support of the Anusandhan National Research Foundation (ANRF) under the Partnerships for Accelerated Innovation and Research (PAIR) project, Government of India, sanction order ANRF/PAIR/2025/000011/PAIR-B. The authors are grateful to the anonymous reviewer and the Editor for the insightful comments to help improve the quality and presentation of this paper. The data was collected from the sites: https://www.nseindia.com/.

Author contributions: AMJ: Conceptualization, Methodology, Software, Formal analysis. NMH: Investigation, Writing - original draft SS: Supervision, Project administration, Writing - review & editing, Funding acquisition AD: Supervision, Writing - review & editing VS: Funding acquisition, Writing - review & editing All authors reviewed the draft and approved the final form of the research.

Funding: This research work is partially supported by the Anusandhan National Research Foundation (ANRF) under the Partnerships for Accelerated Innovation and Research (PAIR) project, Government of India, sanction order ANRF/PAIR/2025/000011/PAIR-B.

AI Disclosure

During the preparation of this work the author(s) used generative AI in order to improve the language of the article. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Agarwal, T., Kumar, S., & Singh, S.P. (2014). Factors affecting movement of Indian stock market: a study with special reference to CNX nifty. *Management Dynamics, 14*(2), 9-18. https://doi.org/10.57198/2583-4932.1099

Ahn, J.M., Kim, J., & Kim, K. (2023). Ensemble machine learning of gradient boosting (XGBoost, LightGBM, CatBoost) and attention-based CNN-LSTM for harmful algal blooms forecasting. *Toxins*, *15*(10), 608. https://doi.org/10.3390/toxins15100608

Ben Jabeur, S., Gharib, C., Mefteh-Wali, S., & Ben Arfi, W. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change, 166*, 120658. https://doi.org/10.1016/j.techfore.2021.120658

Chandrika, G.N., Gumudavelli, S.V.R., Kalleru, A., Kambhampati, V., & Kandaggatlta, P. (2023). Comparative analysis of machine learning algorithms to forecast Indian stock market. *ITM Web of Conferences, 56*, 05009. https://doi.org/10.1051/itmconf/20235605009

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of Xtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing, 73*, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029.

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., & Fang, X. (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access, 7*, 150960-150968. https://doi.org/10.1109/access.2019.2946980

Dong, J., Zeng, W., Wu, L., Huang, J., Gaiser, T., & Srivastava, A.K. (2023). Enhancing short-term forecasting of daily precipitation using numerical weather prediction bias correcting with XGBoost in different regions of China. *Engineering Applications of Artificial Intelligence, 117*(Part A), 105579. https://doi.org/10.1016/j.engappai.2022.105579

Espegren, K., Damman, S., Pisciella, P., Graabak, I., & Tomasgard, A. (2021). The role of hydrogen in the transition from a petroleum economy to a low-carbon society. *International Journal of Hydrogen Energy, 46*(45), 23125-23138. https://doi.org/10.1016/j.ijhydene.2021.04.143

Gülmez, B. (2023). Stock price prediction with optimized deep LSTM network with artificial rabbits optimization algorithm. *Expert Systems with Applications, 227,* 120346. https://doi.org/10.1016/j.eswa.2023.120346

Hartanto, A., Kholik, Y.N., & Pristyanto, Y. (2023). Stock price time series data forecasting using the light gradient boosting machine (LightGBM) model. *International Journal on Informatics Visualization, 7,* 2270-2279. https://doi.org/10.30630/joiv.7.4.01740

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

Jarrah, M., & Derbali, M. (2023). Predicting Saudi stock market index by using multivariate time series based on deep learning. *Applied Sciences, 13*(14), 8356. https://doi.org/10.3390/app13148356

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y (2017). LightGBM: a highly efficient gradient boosting decision tree. In *31st Conference on Neural Information Processing Systems* (pp. 3147-3155). Long Beach, CA, USA.

Kovač, A., Paranos, M., & Marciuš, D. (2021). Hydrogen in energy transition: a review. *International Journal of Hydrogen Energy, 46*(16), 10016-10035. https://doi.org/10.1016/j.ijhydene.2020.11.256

Lee, D.-H., Hsu, S.-S., Tso, C.-T., Su, A., & Lee, D.-J. (2009). An economy-wide analysis of hydrogen economy in Taiwan. *Renewable Energy, 34*(8), 1947-1954. https://doi.org/10.1016/j.renene.2008.12.006

Ma, J., & Liu, L. (2008). Multivariate nonlinear analysis and prediction of Shanghai stock market. *Discrete Dynamics in Nature and Society, 2008*, 526734. https://doi.org/10.1155/2008/526734

Mehtab, S., & Sen, J. (2020). A time series analysis-based stock price prediction using machine learning and deep learning models. *Statistical Finance.* https://doi.org/10.48550/arXiv.2004.11697

Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications, 4*(2), 13-29. https://doi.org/10.5121/ijcsea.2014.4202

Nakagawa, K., & Yoshida, K. (2022). Time-series gradient boosting tree for stock price prediction. *International Journal of Data Mining, Modelling and Management, 14*(2), 110-125. https://doi.org/10.1504/IJDMMM.2022.123357

Pareek, A., Dom, R., Gupta, J., Chandran, J., Adepu, V., & Borse, P.H. (2020). Insights into renewable hydrogen energy: recent advances and prospects. *Materials Science for Energy Technologies, 3*, 319-327. https://doi.org/10.1016/j.mset.2019.12.002

Rezaei, H., Faaljou, H., & Mansourfar, G. (2021). Stock price prediction using deep learning and frequency decomposition. *Expert Systems with Applications, 169*, 114332. https://doi.org/10.1016/j.eswa.2020.114332

Shah, D., Isah, H., & Zulkernine, F. (2019). Stock market analysis: a review and taxonomy of prediction techniques. *International Journal of Financial Studies, 7*(2), 26. https://doi.org/10.3390/ijfs7020026

Sharma, S., Agarwal, S., & Jain, A. (2021). Significance of hydrogen as economic and environmentally friendly fuel. *Energies, 14*(21), 7389. https://doi.org/10.3390/en14217389

Shrivastav, L.K., & Kumar, R. (2022). Gradient boosting machine and deep learning approach in big data analysis: a case study of the stock market. *Journal of Information Technology Research, 15*(1), 1-20. https://doi.org/10.4018/JITR.2022010101

Sirisha, U.M., Belavagi, M.C., & Attigeri, G. (2022). Profit prediction using ARIMA, SARIMA, and LSTM models in time series forecasting. *IEEE Access, 10*, 124715-124727. https://doi.org/10.1109/access.2022.3224938

Su, J., Wang, Y., Niu, X., Sha, S., & Yu, J. (2022). Prediction of ground surface settlement by shield tunneling using XGBoost and Bayesian optimization. *Engineering Applications of Artificial Intelligence, 114*, 105020. https://doi.org/10.1016/j.engappai.2022.105020

Sun, Y., & Tian, L. (2023). Research on stock prediction based on LSTM and CatBoost algorithm. In *Proceedings of the 2nd International Conference on Bigdata Blockchain and Economy Management.* EAI. Hangzhou, China. https://doi.org/10.4108/eai.19-5-2023.2334326

Yadav, A., Jha, C.K., & Sharan, A. (2020). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science, 167*, 2091-2100. https://doi.org/10.1016/j.procs.2020.03.257

Yuan, G., Zhang, T., Zhang, W., & Li, H. (2021). Analysis of stock price based on the XGBoost algorithm with EMA-19 and SMA-15 features. In *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering* (pp. 1-4). IEEE. SC, USA. https://doi.org/10.1109/CSAIEE54046.2021.9543136

Zhou, F., Zhang, Q., Sornette, D., & Jiang, L. (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Applied Soft Computing, 84*, 105747. https://doi.org/10.1016/j.asoc.2019.105747