IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 June 2025, accepted 4 July 2025, date of publication 15 July 2025, date of current version 24 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3589063

RESEARCH ARTICLE

A Multi-Modal Approach Using a Hybrid Vision Transformer and Temporal Fusion Transformer Model for Stock Price Movement Classification

IBANGA KPEREOBONG FRIDAY, (Member, IEEE), SARADA PRASANNA PATI, AND DEBAHUTI MISHRA, (Senior Member, IEEE)
Department of Computer Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha 751030, India
Corresponding author: Debahuti Mishra (debahutimishra@soa.ac.in)

ABSTRACT Stock market price movement primarily focuses on accurately classifying buy and sell signals, which enables traders to maximize profits with well-timed market entry and exit trading positions. This study presents and implements a multi-modal deep learning approach to classifying stock price movement. Our approach adequately captures potential price reversals or continuations by utilizing two modalities (candlestick chart patterns and historical price data). Specifically, the proposed framework converts the historical data into candlestick charts of 256 × 256-pixel images where both modalities are effectively integrated and processed. A key innovation employed is the application of the histogram of oriented gradients (HOG) to extract relevant descriptors, including the candlestick colour, body-to-wick proportions, and wick size. Concurrently, the vision transformer (ViT) model is used to process the images using an embedded projection and multi-head self-attention to extract salient spatial features into a non-overlapping patch of 16 × 16 pixels, which are treated as input tokens for the model. After which, the temporal fusion transformer (TFT) model processes the historical features, candlestick chart features, and the extracted HOG features via a decision-level (late feature fusion) strategy that concatenates these inputs to predict short-term price movements over different horizons (1 day, 3 days, 7 days, and 10 days ahead). We systematically evaluate the model performance using a time series cross-validation split to demonstrate the proposed model's efficacy and generalization across eight indices (BSE, IXIC, N225, NIFTY-50, NSE-30, NYSE, S&P 500, and SSE). The results demonstrate the superior performance of our multi-modal approach, achieving average accuracy, precision, recall, and matthew correlation coefficient (MCC) of 96.17%, 96.24%, 96.15%, and 0.9367, respectively across all evaluated indices. Furthermore, using a real-time trading simulation, the study assesses the practical implications of different window sizes (5, 10, and 15 days). A paired t-test is also conducted to validate the proposed model against benchmarks statistically. The analysis provides valuable insights into how short and long-term traders can effectively maximize the proposed model, highlighting its adaptability for real-world applications.

INDEX TERMS Candlestick, histogram of oriented gradients (HOG), stock price movement, multi-modal, temporal fusion transformer (TFT), vision transformer (ViT).

I. INTRODUCTION

The financial markets exhibit inherent complexity, characterized by a combination of structured, semi-structured, and unstructured qualitative information [1]. This includes significant price volatility, macroeconomic indicators, governmental policies, interest rates, current events, stock-specific news, and real-time news sentiment [1], [2]. These multifaceted factors introduce substantial layers of uncertainty, rendering accurate predictive tasks exceptionally challenging. A study by Shah et al. [3] highlights the sheer volume of data generated by global markets, reaching

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson.

quintillions of bytes, making extracting actionable insights from such vast information streams difficult. Consequently, the critical task for traders to accurately predict stock price movement remains an active and demanding area of research, particularly given the complexities involved in effectively integrating and modelling substantial trading data from different modalities [4], [5]. This typically involves classifying price actions into buy, hold, and sell signals to inform trading decisions and achieve more accurate market entry and exit positions [5].

The field of stock prediction modelling has evolved significantly, progressing from traditional rule-based strategies to machine learning models and, increasingly, to hybrid architectures. These hybrid models leverage the strengths of multiple individual models, each tailored for specific data handling, thereby enhancing model performance [6]. They typically rely on the traditional single-modal approach using only historical trading data. More recently, image-based methods have emerged that transform trading data or price movements into visual representations to leverage powerful spatial models like convolutional neural networks (CNNs). However, these single-modal approaches often overlook the complementary strengths of combining temporal and visual modalities. Notably, recent research has increasingly incorporated other data modalities, including historical trading data, technical indicators, image data, sentiment data, and alternative sources such as web search trends [7], [8]. For instance, advancements in large language models (LLMs) have spurred the development of sentiment-based models capable of analyzing vast quantities of textual data from financial articles, company earnings reports, and social media to derive meaningful trading signals [9], [51].

Among the various tools employed in financial trading, candlestick patterns are recognized as fundamental and highly valuable visual representations. They contain crucial information regarding price momentum, trend reversals, and overall price action [10]. As highlighted in the study by Ha do Prado et al. [11], candlestick charts, originating in 17th-century Japan, effectively present comprehensive price information, including open, high, low, and close prices. Their visual nature allows traders to readily identify market dynamics that might be less apparent when analyzing numerical data alone. While some studies have explored the use of alternative image types, such as heatmaps and technical indicator maps, or combined image representations for stock price movement prediction [12], [13], these analyses can often be subjective and heavily reliant on individual trader expertise. They may not provide a holistic view of market conditions, potentially leading to inconsistencies in trading decisions. Despite significant progress in applying deep learning techniques to this domain, a considerable portion of existing research primarily utilizes either historical price data or textual sentiment data for model training [14], [15]. This approach may inherently overlook the valuable price action information explicitly captured within candlestick chart patterns, which are ubiquitously employed across various trading platforms.

This study addresses these limitations by introducing and implementing a novel multi-modal deep learning approach that synergistically leverages candlestick chart patterns and historical trading data. Distinct from conventional one-day-ahead predictions, our framework aims to classify price movement across multiple forecasting horizons: 1 day, 3 days, 7 days, and 10 days. The key contributions of this study include:

a) Application of a multi-modal fusion, precisely decision-level fusion technique (late fusion), contrasting with early and mid-level fusion methods. Each modality, historical price data, engineered candlestick features, and image embeddings are processed independently using our approach. This allows for fully preserving their unique representations before subsequent fusion, maximizing the information extracted from each source.
b) The histogram of oriented gradients (HOG) [16] is used to capture and extract critical visual descriptors from candlestick charts. These descriptors include essential price movement signals, such as candlestick candle color (bullish or bearish), body-to-wick proportions, and wick size, providing a richer representation of the chart patterns.
c) Integration of the vision transformer (ViT) model [17] to process the candlestick chart images effectively. It employs a token-based patch-embedding extraction mechanism, enabling the capture of intricate spatial features and contextual visual tokens directly from the image data.
d) The temporal fusion transformer (TFT) [18] processes the fused multimodal data. TFT's architecture, comprising gated residual networks, multi-head self-attention mechanisms, and normalization layers, is well-suited for capturing complex temporal dependencies. Furthermore, the TFT model is specifically adapted to handle the multi-horizon classification of price movements (1 day, 3 days, 7 days, and 10 days ahead).
e) The proposed model's efficacy is validated using a time series cross-validation split, ensuring temporal relevance in the evaluation process. The model performance is assessed using classification and profitability evaluation metrics, demonstrating the financial significance of the predictions under realistic trading scenarios, highlighting the proposed framework's practical applicability and potential economic value.

The study is structured as follows: **Section I** gives a brief introduction, and **Section II** presents the gaps in the study after detailing some of the recent work in the field. **Sections III**, **IV**, and **V** delve into the methodology, results, and

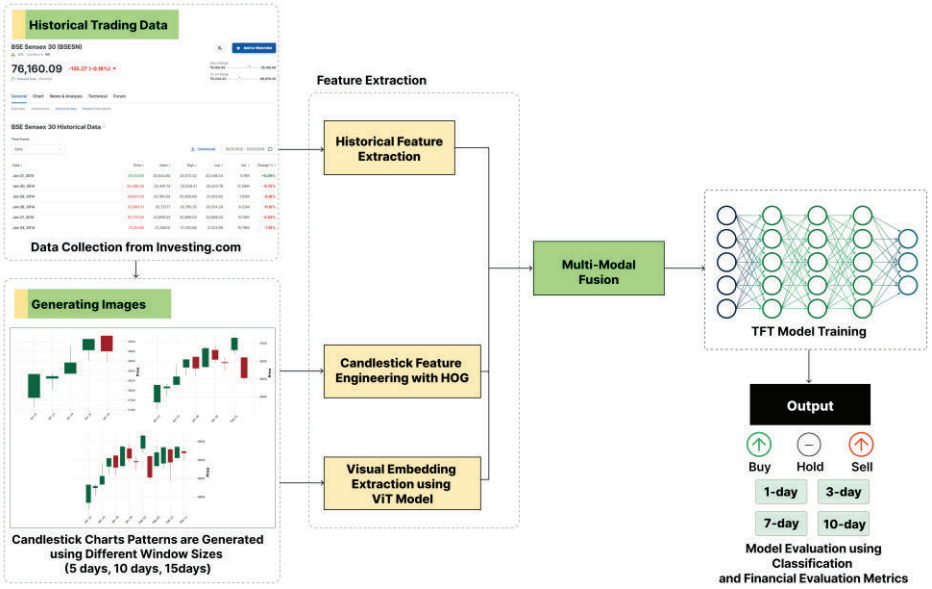

FIGURE 1. Overall proposed framework used in this study.

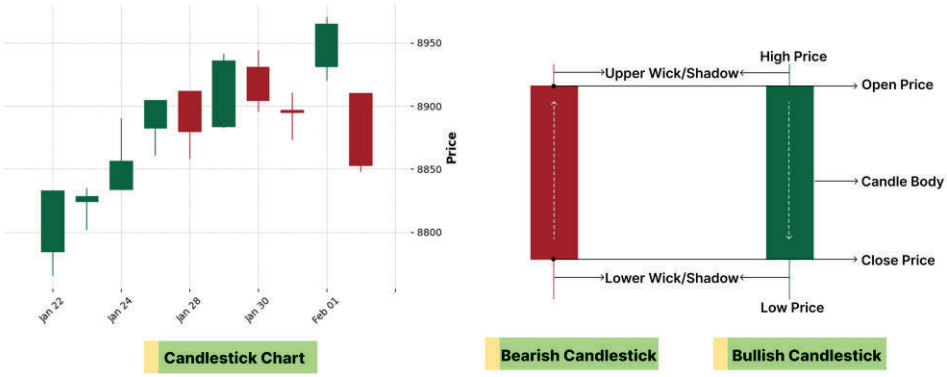

FIGURE 2. Visual representation of a candlestick.

discussion. The study is concluded in **Section VI** with the main findings and contributions.

II. LITERATURE REVIEW

Most research in stock price movement prediction has typically focused on utilizing historical trading data, often augmented with various technical indicators, and employing different model architectures such as gated recurrent units (GRU), long short-term memory (LSTM), and transformer-based models [52]. Advancements in the field have seen the application of image-based stock prediction, notably with the use of CNNs, capitalizing on their success in computer vision and image analysis across various domains. This shift led to studies such as those by Sezer and Ozbayoglu [19], [20], where one-dimensional financial technical analysis data was transformed into two-dimensional images using a 30-day sliding window and subsequently analyzed with a 2D deep CNN for price movement classification. Parallel to the evolution of deep learning model architecture, various studies have introduced innovative approaches employing alternative image representations, including heatmaps, technical indicator maps, and combined image formats. However, while many studies have incorporated multiple data sources, the primary focus has often been on sentiment analysis derived from historical and textual data, with other research integrating textual and candlestick data.

For instance, Zhou et al. [21] investigated the classification of price movements for China A-share market stocks by integrating data from multiple heterogeneous sources, including stock posts, Baidu index data, news articles, and technical indicators, forecasting across three different time horizons: 1 day, 2 days, and 3 days. They emphasized the performance improvements achieved by combining these various data sources and their impact on longer prediction horizons. Additionally, they noted that the effectiveness of their predictions was positively correlated with the trading

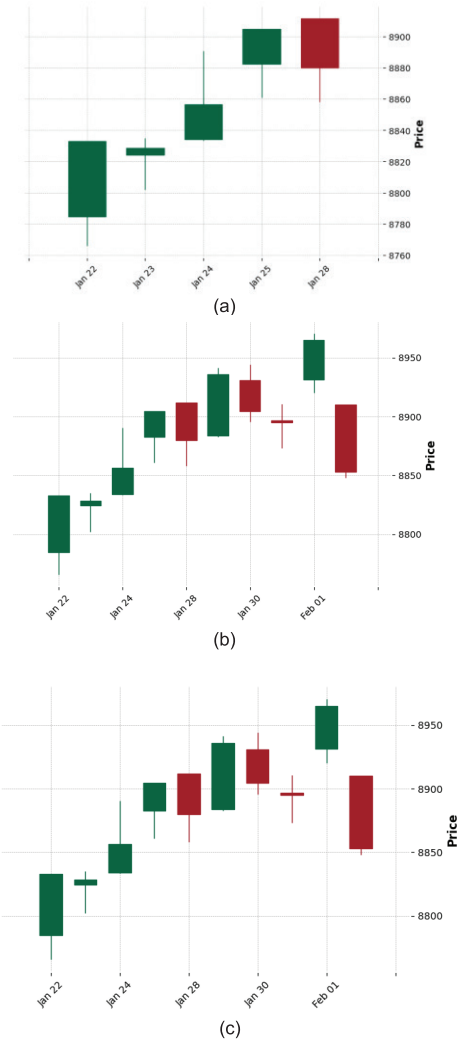

FIGURE 3. Samples of generated candlestick charts for (a) 5 window size, (b) 10 window size, (c) 15 window size.

activity of the stock. However, they did not incorporate any image data within their set of input modalities. In another approach, Shahbandari et al. [22] developed a deep convolutional model where time-series data was transformed into a 2D matrix using gramian angular fields (GAF). They employed the YOLO-v1 model for both the classification and localization of various critical candlestick price patterns, such as morning star, evening star, bullish engulfing, bearish engulfing, shooting star, inverted hammer, bullish harami, and bearish harami. Their research focused on object detection within these transformed images, aiming to identify and classify specific candlestick patterns relevant for informing market entry or exit trading decisions based on the recognized pattern. Similarly, Carta et al. [23] adopted an approach akin to [22], where meta-features were extracted from GAF images generated from historical trading data. These meta-features were subsequently utilized in their proposed stacking ensemble method, which employed majority voting in the final prediction layer. They analyzed historical trading data for the S&P500 and MSFT at various time resolutions, including 5-minute, 10-minute, and 1-hour intervals utilizing a deep double Q reinforcement learning model that fused different signals from training iterations of a meta-learner to process trading signals, employing a reward-based classifier for the final output. Their findings indicated that the meta-learner yielded improved results, exhibited greater resilience to overfitting, and outperformed the ensemble baseline models considered. Furthermore, Lin et al. [24] introduced an eight-trigram classification system for two-day k-line patterns, derived from applying principles of Taoist cosmology to candlestick charts. Their analysis of various candlestick patterns identified those with more significant predictive power, leveraging an ensemble learning framework comprising six distinct models: random forest, gradient-boosted decision trees, logistic regression, k-nearest neighbors, support vector machine, and LSTM. Notably, their experimental evaluation also included an investment strategy based on the model's predictions, demonstrating a higher sharpe ratio (SR) and a lower maximum drawdown (MDD) compared to the buy-and-hold method.

In another intriguing study focusing on candlestick images, Tsai and Quan [25] employed wavelet-based texture feature extraction, deviating from the conventional use of trading data. Their approach applied a three-scale haar wavelet transform (HWT) to the candlestick charts and extracted features from the third-level low-pass sub-band (LL3), which was identified as containing relevant information. These texture-based features reflected structural attributes of the candlestick elements, including the candle body and upper and lower shadows, effectively representing market volatility and directional movements over a specified number of days. This methodology aims to preserve the stock data's visual and spectral characteristics within an image space. Complementarily, Liu et al. [26] presented a multi-type data fusion framework, employing a deep learning module for feature abstraction and a reinforcement learning model for policy optimization and adaptive decision-making. Their study integrated numerical trading data and candlestick chart images. The CNN and bidirectional LSTM models were used for initial and temporal feature extraction from the candlestick images, while an LSTM model processed the trading data. The fused features were used as environmental states for a proposed dueling deep Q-network. This network enabled the learning of trading policies by decoupling the estimation of state-value and advantage functions. This led to more stable learning for buy/sell decisions through exploration and exploitation over time. Their study utilized a feature-level fusion method, where the extracted features from disparate data sources were combined early into a comprehensive feature vector before being processed by the reinforcement learning module. Lim et al. [18] demonstrated the TFT model's application for multi-horizon prediction across various time-series datasets, including electricity, traffic, and retail, highlighting its efficiency for such a prediction task. Although it was not applied to stock data, this makes the TFT model a relevant

model architecture for consideration in this work. In parallel, Gezici and Sefer [27] explored multiple transformer-based architectures—including the ViT, data-efficient image transformer (DeiT), and swin transformer for stock asset trend classification. Their approach involved transforming one-dimensional historical stock data into two-dimensional representations using over 60 technical indicators. These transformed financial images were then fed into the models to predict asset movements as buy, hold, or sell using threshold-driven labeling. The ViT model outperformed DeiT, Swin, ConvMixer, and a CNN-TA++ baseline across predictive and financial performance metrics. Similarly, Tuner et al. [28] also introduced two transformer-based models -DAPP (deep attention-based price prediction), based on Vision Transformers, and DPPP (deep patch-based price prediction), based on convolutional patch embedding to predict asset prices and directional movements using transformed image-like representations of historical time-series data.

Despite working with a relatively small and imbalanced dataset (using a threshold of 0.01 for labeling buy, hold, and sell), both models achieved similar classification performance, with overall accuracy around 62–63%. The study emphasized that model performance limitations in DAPP were attributed to data scarcity, suggesting that transformer-based models may deliver better performance with larger datasets.

Recent research across various fields highlights the power of ViT models, both standalone and integrated with temporal architectures like the TFT. In medical diagnostics, ViT has enhanced physiological signal analysis and image interpretation. For example, by transforming electroencephalogram signals into topographical images, ViT has been shown to improve seizure detection and mental state classification [29], excelling at capturing long-range spatial dependencies. Similarly, in energy forecasting, temporal energy usage patterns can be predicted using the TFT model as in the study by Nazir et al. [30].

These successes demonstrate the unique advantage of both models in their ability to extract spatial structure from visual data and learn complex temporal dependencies via attention mechanisms. Drawing on this proven versatility, our current work applies this philosophy to financial prediction. We integrate ViT-derived visual embeddings from candlestick charts with temporal features modeled by TFT to leverage both high-dimensional spatial patterns and sequential price dynamics for multi-horizon stock movement classification.

In contrast to prior works that employed either multi-modal or single-modal approaches with candlestick images and trading data, which typically relied on low-level visual texture features or HWT coefficients, this study addresses these limitations by leveraging both high-level embeddings and engineered candlestick features. This allows for the exploitation of both local and global patterns inherent within the candlestick charts, specifically incorporating visual technical interpretations such as wick length, body colour, and candle size. Furthermore, our study includes comprehensive classification and profitability prediction experiments across multiple time horizons. It also evaluates the impact of the historical window size on model performance, providing a more thorough understanding of the model's abilities and limitations.

TABLE 1. Hardware and software specifications used for the experiments.

Configuration	Specifications
Execution Environment	Google Colab
Python version	Python 3.10.12
CPU	Intel ® Xeon ® CPU @ 2.30 GHz
GPU	Tesla T4 NVIDIA-SMI 535.104.05
Disk Space	107.7GB

TABLE 2. Summary statistics of selected indices.

Index	Description	Count	Mean	Std. Dev
BSE	Bombay Stock Exchange	2721	43096.77	16914.80
IXIC	NASDAQ-100 Index	2755	9380.42	4226.04
N225	Nikkei 225 Stock Average	2486	21775.61	5559.57
NIFTY50	National Stock Exchange of India 50	2705	1275.15	5050.98
NSE-30	Nigerian Stock Exchange 30	1901	1867.69	793.00
NYSE	New York Stock Exchange	2755	13424.05	2626.77
S&P500	Standard & Poor 500	2755	3232.97	1110.47
SSE	Shanghai Composite Index	2663	3114.14	430.12

III. METHODOLOGY

This section outlines the proposed methodology, detailing the architecture and components of our multi-modal deep learning framework. The proposed framework leverages the HOG [31] for candlestick-engineered features, ViT [27] for spatial feature extraction, and TFT [32] for temporal sequence modelling and multi-horizon prediction. Figure 1 visually depicts the overall workflow, from data collection to candlestick chart generation to modality-specific preprocessing, joint model training, and performance evaluation.

For candlestick chart generation, a fixed-length window size of N days (where $N = 5$, 10, and 15 days) are used to process the images, similar to the approach used in Kusuma et al. [33], emulating consistent, well-established practices in the field. Each candlestick within the chart visually encodes critical price points: the vertical span of the candle's body represents the range between the opening and closing prices. At the same time, the upper and lower wicks (shadows) indicate the intra-period high and low prices, respectively. The color of the candle body denotes the direction of price movement within the period (e.g., green for bullish clos-

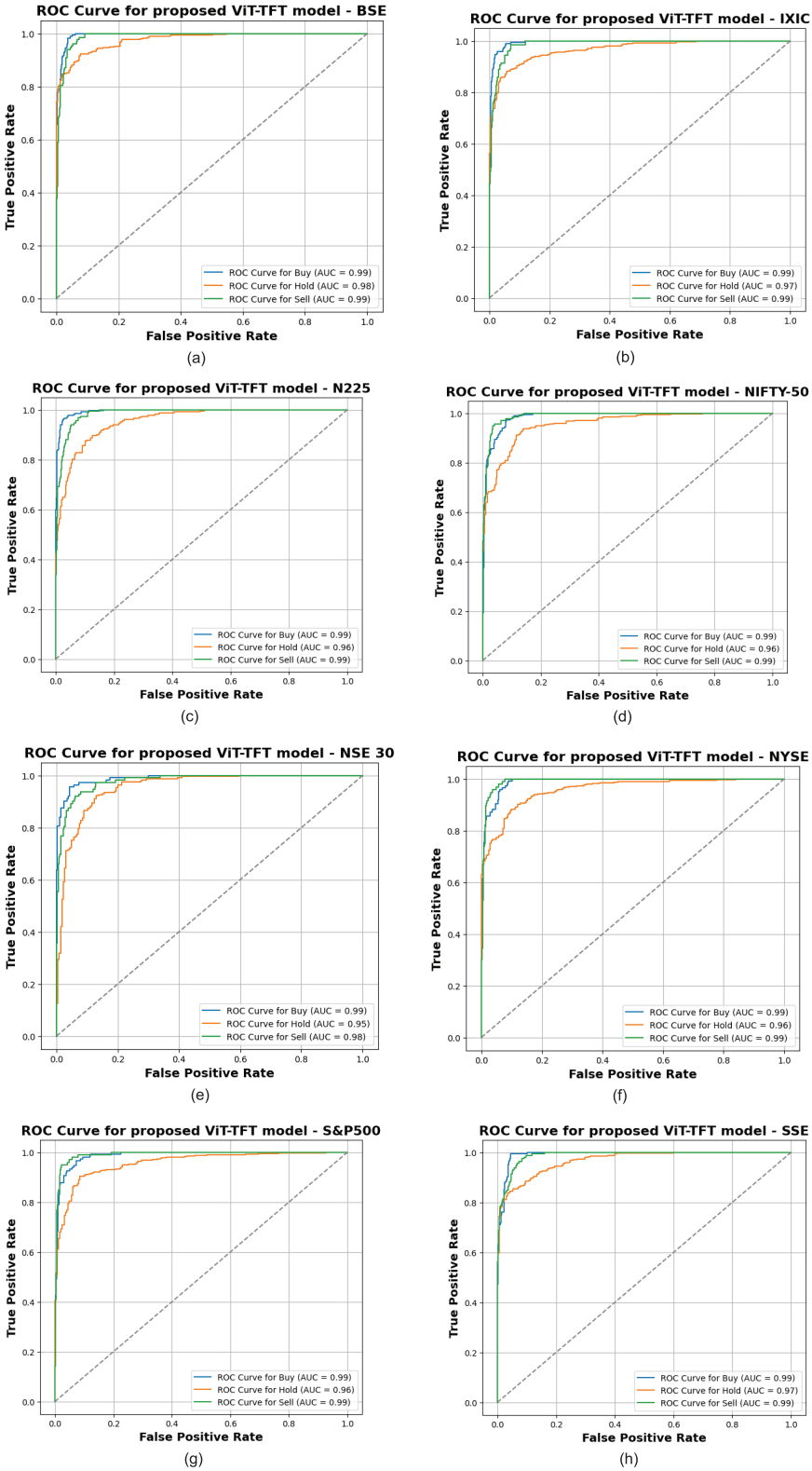

FIGURE 4. ROC Curve for proposed ViT-TFT model using 5 window sizes on (a) BSE (b) IXIC (c) N225 (d) NIFTY50 (e) NSE-30 (f) NYSE (g) S&P500 (h) SSE.

ing above opening, red for bearish closing below opening). **Figure 2** illustrates a pictorial explanation of the features of the candlestick.

A python library, matplotlib finance (mplfinance) [34], is employed to render the charts programmatically. Each chart is generated with a resolution of 256 × 256 pixels, a DPI

TABLE 3. Classification performance of the proposed ViT-TFT model using different input modalities.

	HD		HD + CC		HD + CC + HOG	
	Accuracy	MCC	Accuracy	MCC	Accuracy	MCC
BSE						
1-day	95.45	0.9081	92.13	0.8802	92.71	0.8874
3-day	83.22	0.7594	84.76	0.7725	86.30	0.7926
7-day	80.48	0.7208	83.17	0.7661	88.74	0.8129
10-day	77.36	0.6793	80.01	0.6908	83.05	0.7048
IXIC						
1-day	93.21	0.8993	93.74	0.9030	94.14	0.9123
3-day	84.91	0.7883	86.92	0.7906	88.02	0.8078
7-day	80.44	0.7652	84.35	0.7838	89.67	0.8136
10-day	78.52	0.6754	82.75	0.7011	84.72	0.7166
NIFTY-50						
1-day	90.72	0.8714	92.65	0.8951	94.19	0.9045
3-day	84.70	0.7752	85.32	0.7891	87.33	0.8063
7-day	75.88	0.6625	77.10	0.6738	78.20	0.6920
10-day	72.36	0.6497	75.79	0.6582	80.71	0.6712
S&P500						
1-day	97.12	0.9010	98.35	0.9281	99.78	0.9340
3-day	86.79	0.8082	88.14	0.8223	89.35	0.8381
7-day	83.94	0.7925	86.90	0.8122	90.57	0.8418
10-day	81.38	0.7155	83.30	0.7266	85.33	0.7389

HD – Historical data, CC – candlestick charts, HOG – histogram of gradient features

TABLE 4. Classification performance of the proposed ViT-TFT model using different input modalities.

Window size	Accuracy	Precision	Recall	F1-score	AUC	MCC
BSE						
5	96.42	96.50	96.42	96.43	99.69	0.9414
10	92.71	93.33	92.71	92.71	99.26	0.8874
15	88.87	90.45	88.87	88.37	88.69	0.8225
IXIC						
5	95.75	95.80	95.75	95.71	99.65	0.9367
10	94.14	94.13	94.14	94.10	99.29	0.9123
15	86.26	86.90	86.26	86.46	86.87	0.7932
N225						
5	95.29	95.38	95.29	95.31	99.77	0.9294
10	93.86	93.91	93.86	93.85	98.25	0.9078
15	87.47	89.36	87.47	87.44	88.56	0.8142
NIFTY-50						
5	97.16	97.17	97.16	97.15	99.81	0.9536
10	94.19	94.25	94.19	94.16	99.43	0.9045
15	87.97	89.42	87.97	87.72	88.43	0.8284
NSE-30						
5	95.31	95.36	95.31	95.21	99.65	0.9021
10	92.77	92.93	92.77	92.59	99.39	0.8453
15	85.78	85.87	85.78	85.79	85.90	0.7665
NYSE						
5	96.19	96.18	96.19	96.18	99.77	0.9412
10	95.75	95.99	95.75	95.75	99.69	0.9346
15	85.85	87.54	86.85	86.68	87.53	0.8035
S&P500						
5	96.92	97.04	96.92	96.93	99.91	0.9534
10	95.61	95.79	95.61	95.60	99.78	0.9340
15	84.49	85.82	84.49	84.70	85.93	0.7646
SSE						
5	95.45	95.63	95.45	95.45	99.78	0.9286
10	95.30	95.47	95.30	95.30	99.72	0.9274
15	93.88	93.90	93.88	93.87	89.52	0.9057

(dots per inch) of 80, a JPEG (joint photographic experts group) quality factor of 40, and a classic style to ensure uniform background and high contrast for feature extraction.

Figure 3 depicts sample images of the generated candlestick images across the different window sizes used. During the image generation, we employed the Python garbage collector,

which is explicitly invoked to recover unused RAM resources to maintain computational efficiency and prevent memory-related issues.

A. HISTOGRAM OF ORIENTED GRADIENTS (HOG) FOR CANDLESTICK ANALYSIS

The HOG is a feature descriptor introduced by Dalal and Triggs [16] used for object detection in the field of computer vision by analyzing the local intensity gradients or edge directions within an image. The application of HOG in this study treats the intensity or color variations within and around the candlestick as the basis for gradient calculation. It excels at capturing shape-based and structural cues like edges, candle body geometry, and wick-to-body ratios, which are vital for interpreting price action dynamics. While an alternative like Local Binary Patterns (LBP) [35] is good for texture, HOG's focus on the spatial distribution of edge orientations is more relevant for financial charts, where element orientation directly conveys market sentiment. Also, LBP has rotation invariance that could typically distort the economic meaning of flipped patterns (e.g., a bullish pattern turning into a bearish one), unlike the HOG, which preserves this semantic interpretability.

Each image area is divided into cells, which are further grouped into blocks. Within each cell, orientation bins correspond to the dominant direction of the gradient at the body edges. This process transforms the input image into a feature vector. For an input image, I, the gradient at each pixel (x, y) is computed using first-order derivatives where G_x and G_y represent the gradient magnitudes in the horizontal and vertical directions given in **Equations (1)** and (2), respectively:

$$\frac{\partial I\,(x, y)}{\partial x} \approx G_x = I\,(x+1, y) - i\,(x-1, y) \tag{1}$$

$$\frac{\partial I\,(x, y)}{\partial y} \approx G_x = I\,(x, y+1) - i\,(x, y-1) \tag{2}$$

The gradient magnitude $M\,(x, y)$ and orientation $\theta\,(x, y)$ at each pixel (x, y) are then computed where M represents the strength of the edge, and θ is the direction of the edge, as given in **Equations (3)** and **(4)**.

$$M\,(x, y) = \sqrt{G_x^2 + G_y^2} \tag{3}$$

$$\theta\,(x, y) = tan^{-1}\frac{G_y}{G_x} \tag{4}$$

The HOG descriptor is constructed by dividing the cells into size $c \times c$ pixels. Within each cell, the gradient $\theta\,(x, y)$ are quantized into B orientation bins. For each pixel within a cell, its magnitude $M\,(x, y)$ contributes to the corresponding orientation bin. Let H_b be the accumulated gradient magnitude for bin b, and $I_{bin}\,(\theta\,(x, y)\,, b)$ be an indicator function that ensures only pixels within the corresponding orientation range contribute to the bin, as shown in **Equation (5)**.

$$H_b = M\,(x, y)\,.I_{bin}\,(\theta\,(x, y)\,, b) \tag{5}$$

A local contrast normalization is applied by grouping multiple cells into blocks of size $b \times b$ such that it enhances the method against contrast variations. The HOG feature vector within the block is normalized using the L2-norm, where H is the unnormalized HOG vector for the block, and ϵ is a small positive constant to prevent division by zero as in **Equation (6)**.

$$H_{norm} = \frac{H}{\sqrt{||H||_2^2 + \epsilon}} \tag{6}$$

Once all blocks within the image are processed, their normalized feature vector is concatenated to form the final HOG descriptor as in **Equation (7)**, where d is the dimensionality of the HOG feature vector. The extracted HOG feature vector is subsequently fused with other candlestick attributes, where c is the candlestick body color, and s_b, s_u, s_l represents the body and wick sizes, respectively.

$$f_t = [c, s_b, s_u, s_l, h_c] \in \mathbb{R}^k \tag{7}$$

B. VISION TRANSFORMER (ViT) MODEL

The ViT model is a pre-trained transformer architecture by Dosovitskiy et al. [17] that processes images as a sequence of patches instead of the convolutional operation by CNN. The ViT model was chosen over traditional CNNs because it can capture global dependencies through self-attention. This is particularly suited for image processing, where inter-region relationships (between neighboring candles) are more informative. ViTs treat the entire image as a sequence of patches, making the model learn more effectively across the whole candlestick chart. The input images are divided into a grid of non-overlapping patches, where each patch is then flattened, and the resulting patch vectors are embedded to serve as tokens representing local visual information from different parts of the candlestick image. Positional embeddings are incorporated to retain information regarding the spatial arrangement of these patches. The input candlestick image $I \in R^{H \times W}$ is split into N patches of size $P \times P$, resulting in **Equation (8)**:

$$N = \frac{H \times W}{P^2} \tag{8}$$

Each patch x_i is flattened and mapped into a latent space using a linear projection where $W_e \in \mathbb{R}^{d \times P^2}$ is the learnable embedding matrix, and d is the embedding dimension in **Equation (9)**:

$$z_i = W_e x_i + b_e, i = 1, \ldots, N \tag{9}$$

A learnable class token z_0 is prepended to represent the global feature embedding, where E_{pos} is the positional embedding used to retain spatial information. The feature extraction is done using multi-head self-attention as in **Equation (10)**:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{10}$$

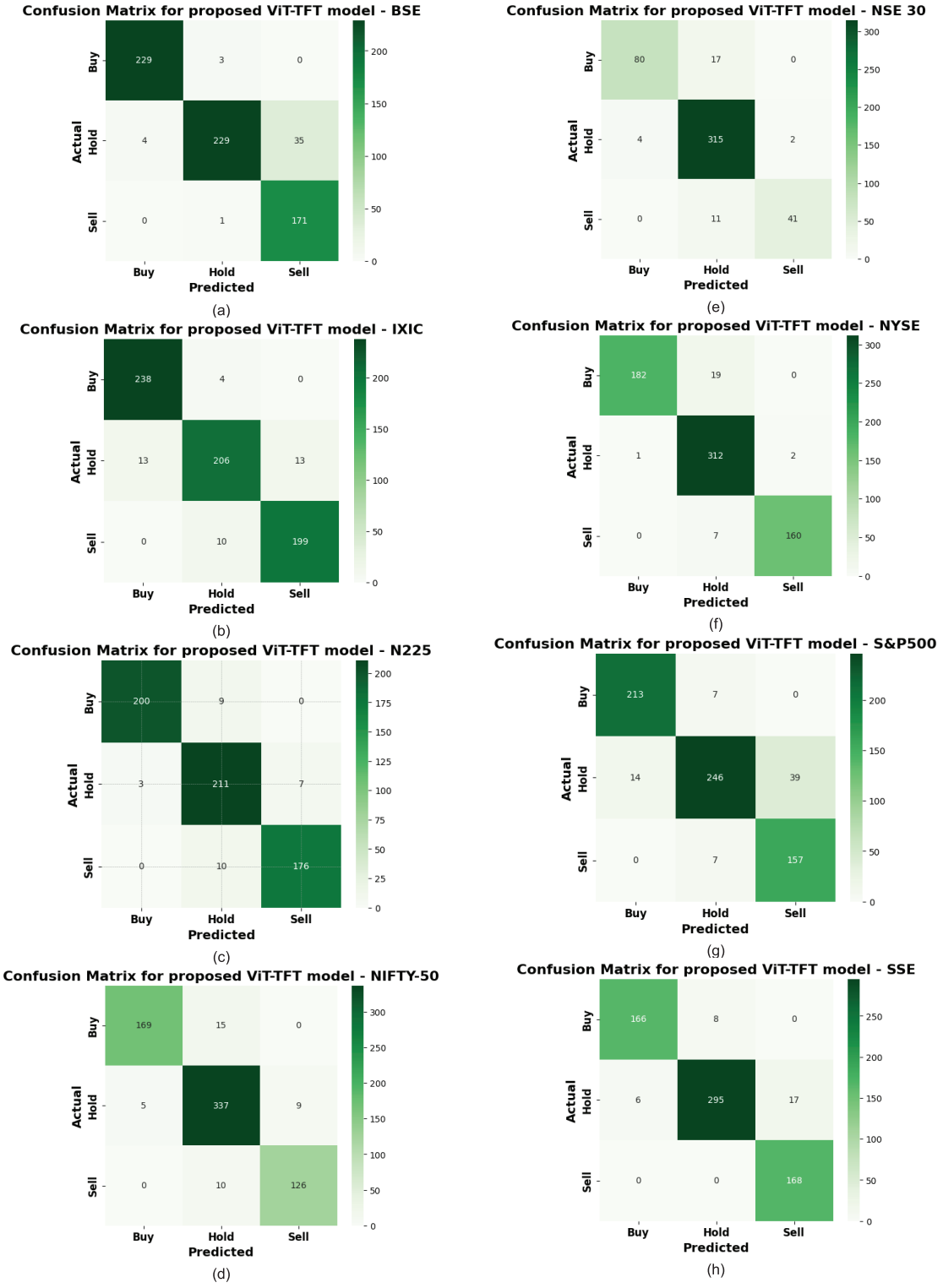

FIGURE 5. Confusion Matrix for proposed ViT-TFT model using 5 window size on (a) BSE (b) IXIC (c) N225 (d) NIFTY50 (e) NSE-30 (f) NYSE (g) S&P500 (h) SSE.

C. TEMPORAL FUSION TRANSFORMER (TFT)

The TFT model by Lim et al. [18] was explicitly designed for multi-horizon time-series forecasting using static covariates, attention, and the gating mechanism that keeps contextual information and prioritizes relevant input variables at each time step for prediction. It integrates static covariates, variable selection networks, and interpretable attention mechanisms to weigh both temporal and static inputs dynamically.

TABLE 5. Classification performance of the proposed ViT-TFT model using different input modalities.

	3 days ahead				7 days ahead				10 days ahead			
	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC	Accuracy	Precision	Recall	MCC
BSE												
5	90.65	90.56	90.47	0.8579	61.07	54.58	55.83	0.3873	44.29	59.27	37.73	0.2162
10	86.30	86.21	87.10	0.7926	88.74	88.13	88.25	0.8129	83.05	83.05	83.05	0.7048
15	82.22	82.04	81.54	0.7310	83.23	85.25	83.92	0.7309	88.18	88.36	88.27	0.7974
IXIC												
5	88.74	88.10	88.06	0.8187	69.72	64.19	66.82	0.4317	66.36	62.40	66.36	0.3777
10	88.02	87.09	86.76	0.8078	89.67	88.91	89.26	0.8136	84.72	78.57	84.72	0.7166
15	85.22	83.89	82.89	0.7629	89.43	88.17	88.17	0.8088	87.94	88.83	87.94	0.7795
NYSE												
5	92.59	90.50	92.48	0.8873	67.92	62.56	62.33	0.4554	66.82	58.85	62.50	0.4156
10	90.24	90.32	90.24	0.8519	92.81	92.63	92.70	0.8821	79.88	77.02	77.54	0.6636
15	90.00	89.84	89.74	0.8477	86.56	86.57	84.53	0.7845	91.91	98.12	91.89	0.8634
NIFTY-50												
5	87.92	88.76	87.19	0.8225	69.31	58.42	63.37	0.4632	47.76	61.63	42.77	0.2497
10	87.33	87.12	87.13	0.8063	78.20	88.79	80.61	0.6920	80.71	80.23	79.18	0.6712
15	85.39	85.50	84.77	0.7805	80.31	87.11	82.00	0.7005	87.23	88.79	87.87	0.7834
NSE-30												
5	84.38	87.22	84.57	0.7713	64.10	65.92	64.18	0.4345	48.65	54.98	46.37	0.2576
10	82.03	84.77	82.21	0.7303	83.30	87.14	84.08	0.7523	78.08	81.48	79.10	0.6544
15	75.88	79.50	76.09	0.6381	81.29	84.19	81.80	0.8177	85.94	85.71	85.74	0.7608
S&P500												
5	90.81	90.67	90.54	0.9833	67.77	58.25	62.54	0.4344	67.27	60.30	63.31	0.3845
10	89.35	89.60	88.80	0.8381	90.57	90.22	90.35	0.8418	85.33	83.05	83.29	0.7389
15	89.25	89.03	88.89	0.8342	88.97	88.25	88.50	0.8144	90.23	89.08	88.94	0.8274
SSE												
5	93.56	93.99	93.65	0.9042	60.56	53.47	54.47	0.3669	59.34	51.56	54.25	0.3261
10	92.34	92.30	92.30	0.8845	92.09	91.90	91.90	0.8758	78.53	76.81	77.52	0.6446
15	89.49	89.43	89.49	0.8415	89.36	88.93	88.97	0.8331	90.35	90.13	90.21	0.8433

In this study, the TFT is employed to process multi-modal stock data using an attention-based feature selection and gated residual network (GRN) to handle short and long-term dependencies in data, which are optimized using the $L2$ regularization for generalization. The study adopts the late fusion (decision-level) method, where the extracted features from the three inputs (historical price data, HOG features, and candlestick features) are concatenated before being fed into the classification layers. This ensures that each feature set retains its independence during the feature extraction while allowing interaction as it is passed through the TFT model, as in **Equation (11)**.

$$X_{TFT} = TFT([x_{hist}, x_{HOG}, x_{VIT}]) \tag{11}$$

The TFT model uses an attention module to focus and capture long-term dependencies in the data, such that for each attention head i, we compute $Q_i = W_i^Q X_{fused}$, $K_i = W_i^K X_{fused}$, and $V_i = W_i^V X_{fused}$, which are learnable weight matrices for queries, keys and values. The attention scores are computed using the scaled dot-product attention, where d_k is the key dimension as given in **Equation (12)**.

$$A_i = softmax \frac{Q_i K_i^T}{\sqrt{d_k}} V_i \tag{12}$$

The final attention output is computed as given in **Equation (13)**, where h is the number of attention heads, and W^O is the learnable output projection matrix.

$$MHA\left(X_{fused}\right) = Concatenate\left(A_1, A_2, ..., A_h\right) W^O \tag{13}$$

$$X_{attn} = X_{fused} + MHA(X_{fused}) \tag{14}$$

The output from the attention module, as in **Equation (14)**, is passed through the GRN, allowing the model to focus on the relevant temporal features where Z_1, Z_2 are linear transformation layers with $ReLU$ activation as given in **Equations**

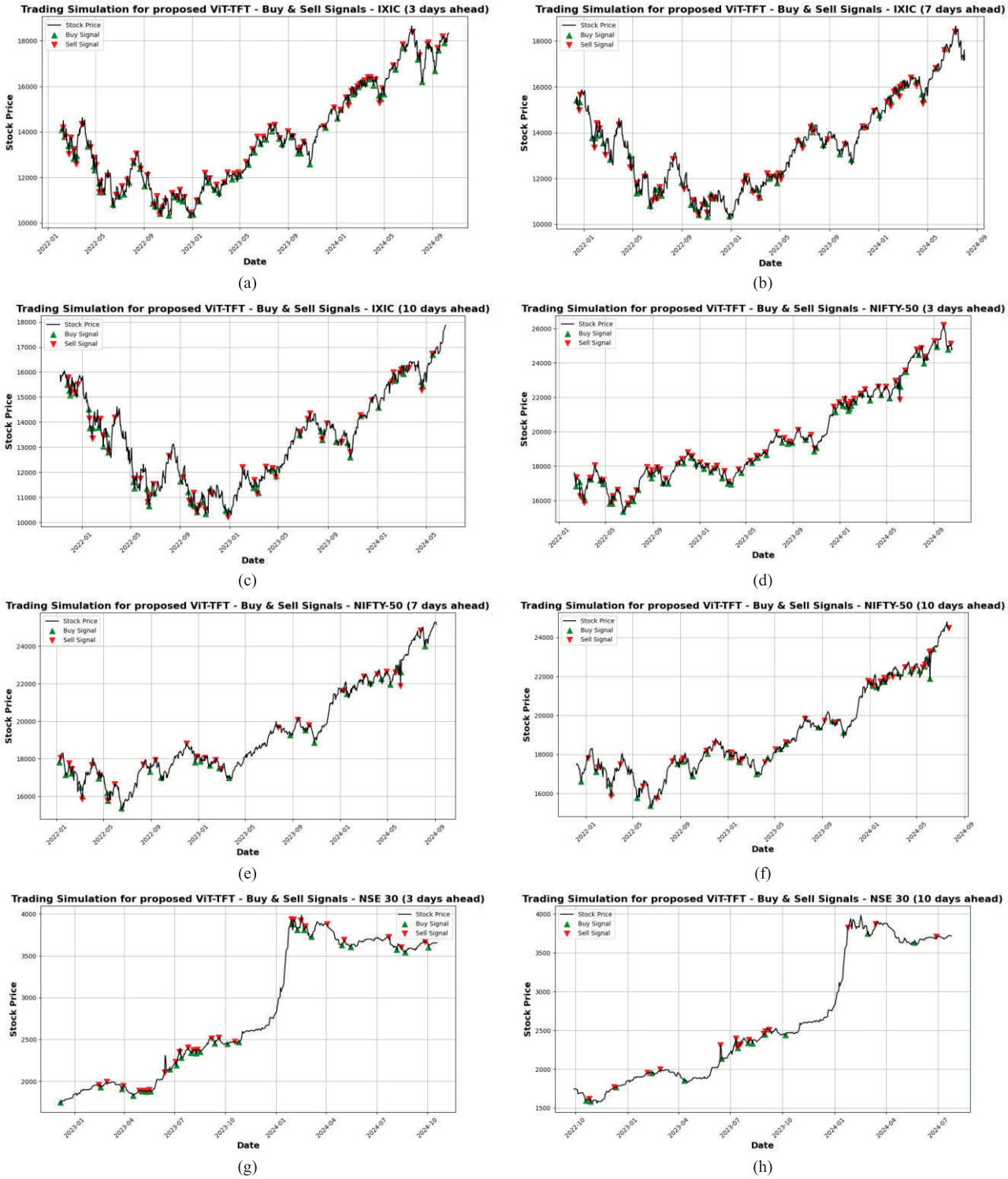

FIGURE 6. Trading buy and sell signals for the proposed ViT-TFT model for 3 days ahead, 7 days ahead, and 10 days ahead across different indices.

(15) to (17), which are passed through the final classification head consisting of dense layers as in **Equations** (18) to (20).

$$Z_1 = ReLU(W_1 X_{attn} + b_1) \tag{15}$$

$$Z_2 = (W_2 Z_1 + b_2) \tag{16}$$

$$X_{GRN} = LayerNorm(Z_2 + X_{attn}) \tag{17}$$

$$X_{flatten} = Flatten(X_{GRN}) \tag{18}$$

$$X_{dense} = ReLU(W_d X_{flatten} + b_d) \tag{19}$$

$$Y_{pred} = softmax(W_{out} X_{dropout} + b_{out}) \tag{20}$$

IV. EXPERIMENTAL SETUP AND DATA PREPARATION

All the experiments conducted in this study were implemented on the google colab platform with the following hardware and software specifications outlined in **Table 1**. The subsequent subsections detail the problem definition, data description, data labeling, and evaluation metrics.

A. PROBLEM STATEMENT

This study aims to classify stock price movements using a multi-modal framework that integrates numerical, visual, and engineered features from candlestick images. The numerical features include historical price data and quantitative extraction of shape-related details using histogram analysis and edge detection. The visual features I_t are represented by the RGB image of the generated candlestick chart given as $I(t) = f_{mpf}(X_t), I(t) \in \mathbb{R}^{256\times256\times3}$, using a rendering function f_{mpf}.

For each time t, a fixed-length window size of $w = (5, 10, 15)$ is applied to generate a candlestick representation of price movement. These features are combined using a joint fusion strategy implemented at the decision level (late fusion). The numerical and engineered features are designed to capture the structural details and quantitative attributes of the candlestick patterns. At the same time, a self-attention mechanism within the TFT model is employed to learn the relative importance and temporal dependencies across the fused multi-modal features for predicting price movements over different horizons.

B. DATA COLLECTION

This study used eight indices, including BSE, IXIC, N225, NIFTY50, NSE, NYSE, S&P 500, and SSE, for the experiments. All the datasets were collected from Investing.com.[1] Each dataset was collected for a period of 10 years, from January 1, 2014, to January 1, 2025, and contained essential trading information, including close price, open price, high price, low price, volume, and price percentage change. The summary statistics of all the datasets are described in **Table 2**. This initial feature set is augmented with technical indicators related to momentum, volatility, and trend. These include moving average (MA), exponential moving average (EMA), volatility index (VIX), relative strength index (RSI), average true range (ATR), bollinger bands (BB), moving average convergence divergence (MACD), TRIX, and commodity channel index (CCI) [38].

C. DATA LABELING

For the data labeling, we employ a similar methodology to the studies by Zhao and Yang [36] and Friday et al. [37], where the price movement is labeled based on the relative change between the current close price and the close price of the $n-day$ ahead. This approach, as opposed to simply comparing consecutive close prices, which is typically employed, provides a more consistent representation of price trends.

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

TABLE 6. Profitability analysis of the proposed ViT-TFT model using different window sizes across all indices.

Window size	ROI	AR	MDD	Sharpe ratio
BSE				
5	119.20	33.67	17.48	1.83
10	125.42	34.98	19.23	1.79
15	102.43	30.14	14.62	2.02
IXIC				
5	49.15	15.92	7.86	1.94
10	48.87	15.77	8.29	1.90
15	49.33	16.07	7.42	1.95
NIFTY-50				
5	36.48	12.22	9.87	1.85
10	36.73	12.17	10.62	1.82
15	37.59	12.58	9.42	1.89
NSE30				
5	3.48	1.83	2.87	1.45
10	3.31	1.74	2.94	1.48
15	3.37	1.78	2.84	1.46
NYSE				
5	34.98	11.72	10.12	2.03
10	33.46	11.20	10.87	2.02
15	34.93	11.78	9.56	2.01
S&P500				
5	11.51	4.11	4.02	1.99
10	11.45	4.07	3.94	1.96
15	11.45	4.11	4.01	2.01
SSE				
5	6.15	2.22	3.52	1.94
10	6.30	2.27	3.47	1.91
15	6.13	2.23	3.56	1.98

As given in **Equations (21)** and **(22)**, the price movement is labelled upward if the future price increases by more than 0.005%, downward if it decreases by more than 0.005%, and hold if the price change is within ±0.005%.

$$Return_n = \frac{close_t - close_{t-n}}{close_{t-n}} \tag{21}$$

$$Price\ movement_i = \begin{cases} 1, & if\ Return_n > 0.005 \\ 0, & if\ Return_n > 0.005 \\ 2, & -0.005 \leq Return_n \leq 0.005 \end{cases} \tag{22}$$

D. EVALUATION METRICS

All the considered models are evaluated using classification metrics including accuracy, precision, f1-score, recall, matthew correlation coefficient (MCC), and AUC score [39], [40]. The profitability of the model was also assessed through a trading simulation algorithm adapted from Friday et al. [37] which used financial metrics including annualized return, return on investment (ROI), MDD, and SR [41]. These metrics provide a comprehensive evaluation of the models, considering both their classification and potential application in real-time trading systems as given in **Equations (23) – (31)**.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

FIGURE 7. Trading buy and sell signals for the proposed ViT-TFT model for one-day ahead prediction (a) BSE (b) IXIC (c) N225 (d) NIFTY50 (e) NSE-30 (f) NYSE (g) S&P500 (h) SSE.

$$Precision = \frac{TP}{TP + FP} \tag{24}$$

$$Recall = \frac{TP}{TP + FN} \tag{25}$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{26}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)\,(TP + FN)\,(TN + FP)(TN + FN)}} \tag{27}$$

$$Annualized\ return = \frac{Final\ value}{Initial\ value}^{\frac{1}{n}} - 1 \tag{28}$$

$$ROI = \frac{Net\ profit}{Investment\ cost} \tag{29}$$

$$Sharpe\ ratio = \frac{Average\ return - Risk - free\ rate}{Std.\ Dev\ of\ returns} \tag{30}$$

$$MDD = \frac{Peak\ value - Trough\ value}{Peak\ value} \tag{31}$$

V. RESULTS AND DISCUSSION

This section gives a detailed result analysis and discussion across the different horizons, the impact of the time steps on classification, and the profitability performance of the model.

A. COMPARATIVE ANALYSIS OF UNI-MODAL AND MULTI-MODEL APPROACH

From the results presented in **Table 3**, we examine the multi-modal approach and the impact of each input modality on improving classification performance as measured by accuracy and MCC scores using the 10-day window size.

Across the different indices and time horizons, we notice a consistent increase in both metrics when moving from using only historical data (HD) to combining HD with candlestick charts (CC). Finally, the HOG features are integrated with both modalities, aiming to add more meaningful information to enhance model performance. On the BSE index, accuracy is better using only HD and slightly decreases at the 1-day horizon. However, as the prediction horizon increases, a clear improvement in model performance is evident across different horizons. The results for the IXIC, NIFTY-50, and S&P 500 validate the efficacy of this multi-modal approach, particularly for both short and long-term prediction, suggesting that the model's predictive performance suffers from short-term dependencies derived solely from HD. Still, the inclusion of visual cues helps the model better understand trend continuity and reversal. The multi-modal approach achieves better performance for long-term predictions (7 and 10 days ahead), with accuracy improvements ranging from 80.48% to 88.74%, and similarly significant improvements of approximately 9% and 6.6% observed in the IXIC and S&P 500 indices, respectively.

Both visual information, including the extracted features and the candlestick-engineered features using HOG from the candlesticks, demonstrate improved model performance and better scalability across prediction horizons.

B. CLASSIFICATION PERFORMANCE OF THE PROPOSED MODEL USING DIFFERENT WINDOW SIZES FOR 1, 3, 7, AND 10-DAY AHEAD

The classification results of the proposed model, as presented in **Table 4**, indicate that for a 1-day ahead price movement prediction, the model achieved the highest performance when trained with both modalities using a 5-day window size across all eight indices. The classification accuracy ranged from 95.31% for the NSE index to 97.16% for the NIFTY-50 index. This demonstrates the model's strong ability to effectively discriminate between the buy, hold, and sell classes at this short-term horizon. Similarly, the MCC values, ranging from 0.9021 for the NSE to 0.9536 for the NIFTY-50, corroborate the superior performance observed with the 5-day window.

This improved performance with a shorter sequence length is likely due to the reduced presence of temporal distortions compared to the 10- and 15-day window sizes, potentially allowing the model to capture immediate market trends more effectively. Furthermore, the consistent performance across other indices, including the leading US indices (S&P 500 and IXIC) and the NSE-30 index in the African market, highlights the model's robust generalization capability. The Receiver Operating Characteristic (ROC) curves and confusion matrices for all indices are visualized in **Figures 4** and **5**, illustrating a more granular view of the classification performance.

A naive expectation based on the 1-day ahead results might suggest that the 5-day window size would maintain optimal performance for longer prediction horizons. However, the results presented in **Table 5** reveal a contrasting trend. While the 5-day window yielded the best results for 1-day and 3-day classification, its performance significantly deteriorated for longer horizons (7 and 10 days). This observation suggests that the optimal temporal context required for accurate prediction varies with the forecasting horizon, implying that a longer lookback period becomes necessary for capturing longer-term market dynamics. On the other hand, the model demonstrated consistent performance for the 7-day and 10-day predictions when trained with the 10-day and 15-day window sizes, with accuracy ranging from 78.20% on the NIFTY-50 (with a 10-day window) to 92.81% on the NYSE index. Overall, the 10-day window size appeared to provide a better balance for both short-term and longer-term predictions, maintaining relatively consistent model performance across all indices.

For the 10-day-ahead predictions, the 15-day window size generally yielded the best results, with accuracy ranging from 87.94% for the IXIC to 91.91% for the NYSE. This underscores the model's increasing need for more extensive temporal information and memory to discern and model longer-term market trends effectively.

TABLE 7. Average computational time for each model across different input modalities.

Model	Input Modality	Train Time	Test Time
CNN	HD	102	7
	HD + CC	128	11
	HD + CC + HOG	136	13
LSTM	HD	111	9
	HD + CC	134	13
	HD + CC + HOG	142	15
ViT	HD	195	23
	HD + CC	210	26
	HD + CC + HOG	218	28
TFT	HD	155	18
	HD + CC	175	22
	HD + CC + HOG	186	25
ViT-TFT	HD	202	27
	HD + CC	235	31
	HD + CC + HOG	270	36

This is a key strength of the proposed ViT-TFT model, as its multi-modal approach, incorporating both historical price data and visual patterns from candlestick charts, allows it to leverage broader market context beyond immediate past prices. This is further supported by the findings of Friday et al. [37], where a single-modality approach utilizing only historical trading data achieved a peak prediction accuracy of only 85% on the SSE index for the 10-day ahead forecast. In contrast, our multi-modal approach achieves significantly higher performance on the SSE index for the same 10-day ahead prediction, with an accuracy of 90.35%, a precision of 90.13%, and an MCC of 0.8433.

C. PROFITABILITY ANALYSIS OF THE PROPOSED MODEL

Profitability analysis is crucial to bridge the gap between classification accuracy and real-world financial outcomes. While high accuracy indicates the model's ability to predict price movements correctly, profitability analysis assesses its practical utility in a trading environment by evaluating the returns generated and the associated risks, thereby determining its economic viability.

As discussed in **Section IV**, we evaluated the proposed model using a real-time trading simulation algorithm [37], with an initial capital of 100,000 and no trading costs. The quantified results of the profitability analysis, presented in **Table 6**, indicate that the model performed optimally with longer prediction horizons in terms of risk-adjusted returns. The selection of indices, including the S&P 500, NYSE, SSE, IXIC, N225, and BSE, was deliberate, aimed at covering diverse market states, including bullish, bearish, and range-bound price movements. For the bullish markets, the model consistently generated profitable signals with low MDD, especially at longer prediction horizons such as 15 days. This is evident in the BSE index, where the 5-day window size generated a higher return on investment (ROI). Still, it also incurred increased drawdowns, suggesting higher volatility and potential for significant losses.

Conversely, the 15-day window achieved the highest SR of 2.02 and the lowest MDD of 14.45%, indicating superior risk-adjusted returns by providing a better balance between profitability and risk. This improved risk management can be attributed to the model's ability to make more informed trading decisions over a longer temporal context, thereby potentially avoiding spurious or false entry signals triggered by short-term market noise. Similar stable performances with low drawdowns were evident on the IXIC (approximately 9% to 10% MDD) and NIFTY-50 (approximately 10% to 12% MDD). The moderate SR on the NSE-30 suggests a market characterized by lower inherent risk and potential return compared to the other indices analyzed. For ranging markets, the model demonstrated adaptive behavior by capturing short-term volatility and preserving capital through minimized losses.

On the NYSE, the model demonstrated optimal performance across all evaluated window sizes, consistently exhibiting a SR greater than 2 with considerably low drawdowns. For the bearish indices like the SSE, despite the prevailing downward market trend, the model effectively generated profitable short signals. The SSE index showed MDDs below 4% across all windows, with the 15-day horizon achieving a SR of 1.98, this would be evidence of robustness even under pessimistic market regimes. The S&P 500 and SSE indices also showed moderate to low returns, with a trend of increasing SR and consecutively decreasing MDD as the timeframe lengthened, suggesting the superiority of the 15-day window size for achieving better risk-adjusted returns in these markets.

Generally, a higher ROI is often accompanied by a higher MDD, as is evident in the results for the BSE index. However, the model effectively navigated alternative markets like the NYSE and IXIC, which appeared to offer high returns and low drawdown potential. The choice of window size emerges as a critical factor in determining the profitability risks trade-off. Shorter windows may lead to higher potential ROI but also carry a greater risk of significant drawdowns. In comparison, longer windows tend to result in lower MDD and more stable returns, potentially being more suitable for swing traders who aim to capitalize on medium-term price movements while mitigating short-term volatility. The buy and sell signals generated by the model using the 5-day window for the different prediction horizons are graphically illustrated in **Figures 6 and 7**. The findings of this study are consistent with the work of Gârleanu and Pedersen [50], which demonstrates that optimally designed strategies with strong SR can maintain competitive performance even after accounting for moderate transaction costs, particularly by optimizing trade execution and leveraging persistent trading signals.

D. COMPARISON WITH BENCHMARK STUDIES

To evaluate the proposed model's efficacy, a comparative analysis was conducted against benchmark studies employing

TABLE 8. Performance metrics of the proposed model against state-of-the-art (SOTA) models.

Ref	Year	Modality	Model	Accuracy	F1-score	AUC	MCC
			S&P500				
[42]	2021	CC + HD	2D-CNN	75.38	65.25	62.16	-
[43]	2022	CC	Ensemble XGB	54.00	-	-	-
[44]	2024	CC	YOLOv8 + VGG16	70.10	69.90	69.900	-
[45]	2024	HD	PLSTM-TAL	85.09	86.71	93.12	0.6975
Proposed	**2025**	**HD + CC**	**ViT-TFT**	**96.92**	**96.93**	**99.91**	**0.9534**
			SSE				
[15]	2022	CC	Multiple Attention GNN	66.54	-	-	-
[46]	2023	CS	CS-ACNN	57.30	-	56.80	-
[46]	2023	HD	CS-ACNN	52.20	-	51.70	-
[44]	2024	HD	PLSTM-TAL	88.21	88.80	94.64	0.7637
Proposed	**2025**	**HD + CC**	**ViT-TFT**	**95.45**	**95.45**	**99.78**	**0.9286**
			NIFTY-50				
[47]	2022	CC	CNN-LSTM	57.2	53.1	-	-
[42]	2022	CC	Ensemble XGB	53	-	-	-
[48]	2023	CC	CNN	78	-	-	-
[44]	2024	HD	PLSTM-TAL	85.11	86.76	91.95	0.6988
Proposed	**2025**	**CC + HD**	**ViT-TFT**	**97.16**	**97.15**	**99.81**	**0.9536**

HD – Historical data, CC – Candlestick charts

either candlestick images or historical trading data for stock price movement prediction. This comparison validates the advantages of our multi-modal approach by demonstrating the proposed model's improved classification performance.

As discussed in **Section II**, a significant portion of prior research utilizing candlestick charts has primarily focused on the object detection and classification of specific candlestick patterns, often without directly addressing the prediction of price movements across different future time horizons. Studies relying solely on historical price data have also reported varying levels of accuracy. For instance, the PLSTM-TAL model [45] reported 85.09% and 88.21% classification accuracies for the S&P 500 and SSE indices, respectively, as given in **Table 8**. In another study, the CAGTRADE model [37] achieved average accuracies ranging between 79% and 85% for 7-day and 10-day ahead predictions on indices such as BSE, IXIC, NIFTY-50, and SSE. Notably, the classification accuracies achieved by our proposed multi-modal model, as detailed in **Tables 4** and **5**, consistently surpass these reported benchmark results, ranging from 87% to 91% for comparable prediction horizons and indices. This improved performance highlights the benefit of integrating visual information from candlestick charts with temporal features extracted from historical price data.

E. COMPUTATIONAL COST ANALYSIS

We also evaluated the performance of the proposed model, which showed a high computation demand, considering that the model architecture is a dual-stream data pipeline. Both CNN and LSTM offer low overhead but lack capacity for temporal modelling. Neither model was shown to learn much from either modality. Similarly, the ViT and TFT models exhibited improved performance when handling either historical data alone or images, while using higher training time, as shown in **Table 7**. Both models showed reduced performance when tested for both input modalities. The proposed model took the highest computational time, but adequately showed that it could handle both modalities. This trade-off is justified given the substantial improvement in predictive accuracy during classification and risk-adjusted returns during trading simulation.

F. STATISTICAL VALIDATION

Statistical validation ascertains whether the observed performance improvements of the proposed model over benchmark studies are statistically significant. By employing hypothesis testing, we provide a quantitative measure of confidence in the superiority of the proposed model. The paired t-test [49] was used to statistically validate the superiority of the proposed ViT-TFT model over the best-performing published state-of-the-art benchmarks on three representative indices: NIFTY-50, S&P500, and SSE. The best benchmark model (PLSTM-TAL [45]) for each dataset was selected based on the highest reported average performance, and its results were paired with the corresponding ViT-TFT results across the three datasets, forming three paired samples for each evaluation metric. We tested the following hypotheses using a

TABLE 9. Statistical validation of proposed ViT-TFT for NIFTY-50, S&P500, and SSE.

Metric	ViT-TFT	PLSTM-TAL [45]	*t*-statistic	*p*-value	*h*-value
Accuracy	96.51	86.14	6.616	0.0110	1
AUC	99.83	93.24	8.338	0.0070	1

one-tailed configuration, assuming superiority in the positive direction:

a) Null Hypothesis (H0): There is no significant difference in performance between the ViT-TFT model and the benchmark.
b) Alternative Hypothesis (H1): The ViT-TFT model significantly outperforms the benchmark.

The results summarized in **Table 9** show a t-statistic of 6.616 with a corresponding p-value of 0.0110 for accuracy. This indicates a statistically significant improvement in accuracy achieved by the proposed ViT-TFT model over the PLSTM-TAL benchmark at the 5% significance level (α =0.05). The AUC comparison also returned a t-statistic of 8.338 with a p-value of 0.0070, demonstrating statistically.

significant improvement in the model's ability to accurately distinguish between the buy, sell, and hold classes. These statistical validation results provide strong evidence for the efficacy of the multi-modal approach employed in this study across the selected diverse indices.

G. LIMITATION OF THE STUDY

While this study presents a robust multi-modal framework for stock market prediction, we recognize several inherent limitations that warrant attention and define promising directions for future research.

a. One significant limitation stems from the architectural complexity. The fusion of ViT for image-based features and Temporal Fusion Transformers TFT for historical data, while highly effective in capturing diverse patterns, results in a considerable computational cost. This elevated overhead makes deploying this architecture challenging in resource-constrained environments. Future research will focus on mitigating this by exploring architecture compression techniques, such as pruning and quantization, or by developing more lightweight alternative models specifically designed for efficiency.
b. Furthermore, our current analysis emphasizes risk-adjusted profitability but does not explicitly account for real-world transaction costs. We have not simulated a complete trading strategy that incorporates critical elements like slippage, trading fees, or liquidity constraints. Addressing these practical considerations in future studies is essential for a more accurate assessment of the model's real-world deployability and net profitability.
c. Lastly, the model's performance was not evaluated explicitly against labeled anomaly events. Testing its resilience and predictive power during periods of extreme market volatility or unexpected shocks (e.g., financial crises and geopolitical events) would provide invaluable insights into its robustness. A key future direction involves integrating anomaly detection frameworks or conducting event-driven analyses to understand and improve model behavior under such critical conditions.

VI. CONCLUSION

This study proposes and evaluates a multi-modal deep learning approach for classifying buy, hold, and sell signals for short-term stock market price movements across different horizons: 1, 3, 7, and 10 days. This extends beyond the limitations of prior research, which predominantly focused on one-day-ahead predictions and single-modal approaches. The proposed framework integrates candlestick chart patterns and historical trading data as complementary modalities, processed by applying the HOG, ViT, and TFT models.

The HOG feature descriptor is strategically employed to extract salient candlestick features, including the body-to-wick ratio, wick size, and candle color, effectively capturing essential texture and edge attributes from the visual representations. The ViT model processes the 256 $\times$ 256 pixel candlestick images by embedding non-overlapping 16 $\times$ 16 patches into contextual tokens, extracting relevant global spatial features from the visual modality. The TFT architecture, incorporating gating mechanisms, multi-head self-attention, and normalization layers, serves as the classification module. It integrates the processed features from each modality using a decision-level (late) fusion strategy, ensuring that the unique representations of both visual and temporal data are preserved and independently learned before being combined for prediction.

The proposed model's performance is evaluated using a time series split cross-validation across eight major global stock market indices (BSE, IXIC, NIFTY-50, NSE30, N225, NYSE, S&P500, and SSE). The results demonstrate superior classification performance, achieving average accuracy, precision, recall, and MCC values of 96.17%, 96.24%, 96.15%, and 0.9367, respectively, outperforming previously established benchmark models. Notably, the ViT-TFT model maintained consistent performance across all evaluated prediction horizon lengths, particularly when utilizing a 10-day historical window for both short-term and long-term price movement classification. Furthermore, the practical utility of the model's predictions is validated through a real-time trading simulation, demonstrating its ability to achieve consistently high SR (above 1.8, peaking at 2.02) and low MDD (below 16%), indicating the potential for high risk-adjusted returns in a trading environment. The statistical significance of the model's performance improvement over benchmark models is further confirmed through paired t-tests. This study significantly advances the state-of-the-art multi-modal

market price classification by effectively complementing traditional trading price data with visual and latent cues extracted from candlestick charts.

Future research directions could extend this work by integrating location-sensitive pattern detection modules, enabling the model to identify and assign weights to significant visual motifs and focusing on market actions associated with specific candlestick patterns. Additionally, the model architecture could be expanded to predict multi-horizon outputs simultaneously, improving prediction stability and reducing computational costs by leveraging shared temporal representations across different forecasting windows.

DATA AVAILABILITY

All data used in this study will be made available by the first author upon request.

CONFLICT OF INTEREST

The authors declare no conflict of interest as defined by IEEE, or other interests that might be perceived to influence the results and/or discussion report in this article.

REFERENCES

[1] H. S. Sim, H. I. Kim, and J. J. Ahn, "Is deep learning for image recognition applicable to stock market prediction?" *Complexity*, vol. 2019, no. 1, Jan. 2019, Art. no. 4324878, doi: 10.1155/2019/4324878.

[2] K. Cui, R. Hao, Y. Huang, J. Li, and Y. Song, "A novel convolutional neural networks for stock trading based on DDQN algorithm," *IEEE Access*, vol. 11, pp. 32308–32318, 2023, doi: 10.1109/ACCESS.2023.3259424.

[3] P. Shah, K. Desai, M. Hada, P. Parikh, M. Champaneria, D. Panchal, M. Tanna, and M. Shah, "A comprehensive review on sentiment analysis of social/Web media big data for stock market prediction," *Int. J. Syst. Assurance Eng. Manage.*, vol. 15, no. 6, pp. 2011–2018, Jun. 2024, doi: 10.1007/s13198-023-02214-6.

[4] L. Wang, J. Li, L. Zhao, Z. Kou, X. Wang, X. Zhu, H. Wang, Y. Shen, and L. Chen, "Methods for acquiring and incorporating knowledge into stock price prediction: A survey," 2023, *arXiv:2308.04947*.

[5] M. Wen, P. Li, L. Zhang, and Y. Chen, "Stock market trend prediction using high-order information of time series," *IEEE Access*, vol. 7, pp. 28299–28308, 2019, doi: 10.1109/ACCESS.2019.2901842.

[6] L. Song, S. Chen, Z. Meng, M. Sun, and X. Shang, "FMSA-SC: A fine-grained multimodal sentiment analysis dataset based on stock comment videos," *IEEE Trans. Multimedia*, vol. 26, pp. 7294–7306, 2024, doi: 10.1109/TMM.2024.3363641.

[7] H. Yu, X. Hao, L. Wu, Y. Zhao, and Y. Wang, "Eye in outer space: Satellite imageries of container ports can predict world stock returns," *Humanities Social Sci. Commun.*, vol. 10, no. 1, pp. 1–6, Jul. 2023, doi: 10.1057/s41599-023-01891-9.

[8] K. Obaid and K. Pukthuanthong, "A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news," *J. Financial Econ.*, vol. 144, no. 1, pp. 273–297, Apr. 2022, doi: 10.1016/j.jfineco.2021.06.002.

[9] A. Peivandizadeh, S. Hatami, A. Nakhjavani, L. Khoshsima, M. R. C. Qazani, M. Haleem, and R. Alizadehsani, "Stock market prediction with transductive long short-term memory and social media sentiment analysis," *IEEE Access*, vol. 12, pp. 87110–87130, 2024, doi: 10.1109/ACCESS.2024.3399548.

[10] K. H. Lee and G. S. Jo, "Expert system for predicting stock market timing using a candlestick chart," *Expert Syst. Appl.*, vol. 16, no. 4, pp. 357–364, May 1999, doi: 10.1016/s0957-4174(99)00011-1.

[11] H. A. D. Prado, E. Ferneda, L. C. R. Morais, A. J. B. Luiz, and E. Matsura, "On the effectiveness of candlestick chart analysis for the Brazilian stock market," *Proc. Comput. Sci.*, vol. 22, pp. 1136–1145, Jan. 2013, doi: 10.1016/j.procs.2013.09.200.

[12] A. Brim and N. S. Flann, "Deep reinforcement learning stock market trading, utilizing a CNN with candlestick images," *PLoS ONE*, vol. 17, no. 2, Feb. 2022, Art. no. e0263181, doi: 10.1371/journal.pone.0263181.

[13] P. Khodaee, A. Esfahanipour, and H. Mehtari Taheri, "Forecasting turning points in stock price by applying a novel hybrid CNN-LSTM-ResNet model fed by 2D segmented images," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105464, doi: 10.1016/j.engappai.2022.105464.

[14] L. Jing and Y. Kang, "Automated cryptocurrency trading approach using ensemble deep reinforcement learning: Learn to understand candlesticks," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121373, doi: 10.1016/j.eswa.2023.121373.

[15] J. Wang, X. Li, H. Jia, T. Peng, and J. Tan, "Predicting stock market volatility from candlestick charts: A multiple attention mechanism graph neural network approach," *Math. Problems Eng.*, vol. 2022, pp. 1–16, Sep. 2022, doi: 10.1155/2022/4743643.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, San Diego, CA, USA, Jun. 2005, pp. 886–893, doi: 10.1109/CVPR.2005.177.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[18] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, Oct. 2021, doi: 10.1016/j.ijforecast.2021.03.012.

[19] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," *Appl. Soft Comput.*, vol. 70, pp. 525–538, Sep. 2018, doi: 10.1016/j.asoc.2018.04.024.

[20] O. Berat Sezer and A. Murat Ozbayoglu, "Financial trading model with stock bar chart image time series with deep convolutional neural networks," 2019, *arXiv:1903.04610*.

[21] Z. Zhou, M. Gao, Q. Liu, and H. Xiao, "Forecasting stock price movements with multiple data sources: Evidence from stock market in China," *Phys. A, Stat. Mech. Appl.*, vol. 542, Mar. 2020, Art. no. 123389, doi: 10.1016/j.physa.2019.123389.

[22] L. Shahbandari, E. Moradi, and M. Manthouri, "Stock price prediction using multi-faceted information based on deep recurrent neural networks," 2024, *arXiv:2411.19766*.

[23] S. Carta, A. Corriga, A. Ferreira, A. S. Podda, and D. R. Recupero, "A multi-layer and multi-ensemble stock trader using deep learning and deep reinforcement learning," *Appl. Intell.*, vol. 51, no. 2, pp. 889–905, Feb. 2021, doi: 10.1007/s10489-020-01839-5.

[24] Y. Lin, S. Liu, H. Yang, and H. Wu, "Stock trend prediction using candlestick charting and ensemble machine learning techniques with a novelty feature engineering scheme," *IEEE Access*, vol. 9, pp. 101433–101446, 2021, doi: 10.1109/ACCESS.2021.3096825.

[25] C.-F. Tsai and Z.-Y. Quan, "Stock prediction by searching for similarities in candlestick charts," *ACM Trans. Manage. Inf. Syst.*, vol. 5, no. 2, pp. 1–21, Jul. 2014, doi: 10.1145/2591672.

[26] P. Liu, Y. Zhang, F. Bao, X. Yao, and C. Zhang, "Multi-type data fusion framework based on deep reinforcement learning for algorithmic trading," *Appl. Intell.*, vol. 53, no. 2, pp. 1683–1706, Jan. 2023, doi: 10.1007/s10489-022-03321-w.

[27] A. H. B. Gezici and E. Sefer, "Deep transformer-based asset price and direction prediction," *IEEE Access*, vol. 12, pp. 24164–24178, 2024, doi: 10.1109/ACCESS.2024.3358452.

[28] T. Tuncer, U. Kaya, E. Sefer, O. Alacam, and T. Hoser, "Asset price and direction prediction via deep 2D transformer and convolutional neural networks," in *Proc. 3rd ACM Int. Conf. AI Finance*, Nov. 2022, pp. 79–86.

[29] O. D'Angelis, L. Bacco, L. Vollero, and M. Merone, "Advancing ECG biometrics through vision transformers: A confidence-driven approach," *IEEE Access*, vol. 11, pp. 140710–140721, 2023.

[30] A. Nazir, A. K. Shaikh, A. S. Shah, and A. Khalil, "Forecasting energy consumption demand of customers in smart grid using temporal fusion transformer (TFT)," *Results Eng.*, vol. 17, Mar. 2023, Art. no. 100888, doi: 10.1016/j.rineng.2023.100888.

[31] X. Teng, X. Zhang, and Z. Luo, "Multi-scale local cues and hierarchical attention-based LSTM for stock price trend prediction," *Neurocomputing*, vol. 505, pp. 92–100, Sep. 2022.

[32] X. Hu, "Stock price prediction based on temporal fusion transformer," in *Proc. 3rd Int. Conf. Mach. Learn., Big Data Bus. Intell. (MLBDBI)*, Taiyuan, China, Dec. 2021, pp. 60–66, doi: 10.1109/MLBDBI54094.2021.00019.

[33] R. Mangir Irawan Kusuma, T.-T. Ho, W.-C. Kao, Y.-Y. Ou, and K.-L. Hua, "Using deep learning neural networks and candlestick chart representation to predict stock market," 2019, *arXiv:1903.12258*.

[34] N. S. Raju, J. R. Kumar, and B. Sujatha, "Time series analysis of stock price movements: Insights from data mining using machine learning," in *Proc. AIP Conf.*, vol. 2492, 2023, pp. 1–11, doi: 10.1063/5.0117417.

[35] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proc. 8th Eur. Conf. Comput. Vis.*, Prague, Czech Republic, 2004, pp. 469–481.

[36] Y. Zhao and G. Yang, "Deep learning-based integrated framework for stock price movement prediction," *Appl. Soft Comput.*, vol. 133, Jan. 2023, Art. no. 109921, doi: 10.1016/j.asoc.2022.109921.

[37] I. K. Friday, S. P. Pati, D. Mishra, P. K. Mallick, and S. Kumar, "CAGTRADE: Predicting stock market price movement with a CNN-attention-GRU model," *Asia–Pacific Financial Markets*, vol. 32, no. 2, pp. 583–608, Jun. 2025, doi: 10.1007/s10690-024-09463-w.

[38] A. K. Das, D. Mishra, K. Das, and K. C. Mishra, "A feature ensemble framework for stock market forecasting using technical analysis and Aquila optimizer," *IEEE Access*, vol. 12, pp. 187899–187918, 2024, doi: 10.1109/ACCESS.2024.3461792.

[39] Q. Zhang, Y. Zhang, X. Yao, S. Li, C. Zhang, and P. Liu, "A dynamic attributes-driven graph attention network modeling on behavioral finance for stock prediction," *ACM Trans. Knowl. Discovery Data*, vol. 18, no. 1, pp. 1–29, Jan. 2024.

[40] S. Dash, P. K. Sahu, and D. Mishra, "Forex market directional trends forecasting with bidirectional-LSTM and enhanced DeepSense network using all member-based optimizer," *Intell. Decis. Technol.*, vol. 17, no. 4, pp. 1351–1382, Nov. 2023.

[41] F. Jeribi, R. J. Martin, R. Mittal, H. Jari, A. H. Alhazmi, V. Malik, S. L. Swapna, S. B. Goyal, M. Kumar, and S. V. Singh, "A deep learning based expert framework for portfolio prediction and forecasting," *IEEE Access*, vol. 12, pp. 103810–103829, 2024, doi: 10.1109/ACCESS.2024.3434528.

[42] T.-T. Ho and Y. Huang, "Stock price movement prediction using sentiment analysis and CandleStick chart representation," *Sensors*, vol. 21, no. 23, p. 7957, Nov. 2021, doi: 10.3390/s21237957.

[43] Y. Santur, "Candlestick chart based trading system using ensemble learning for financial assets," *Sigma J. Eng. Natural Sci.*, vol. 40, no. 2, pp. 370–379, 2022, doi: 10.14744/sigma.2022.00039.

[44] N. T. Duong, K. T. Hoang, K. Q. Duong, D. Q. Dinh, H. D. Le, T. H. Nguyen, B. X. Duong, B. Q. Tran, and A. N. Bui, "Investigating market strength prediction with CNNs on candlestick chart images," in *Proc. 6th Asia Conf. Mach. Learn. Comput.*, Jul. 2024, pp. 120–127, doi: 10.1145/3690771.3690776.

[45] S. Latif, N. Javaid, F. Aslam, A. Aldegheishem, N. Alrajeh, and S. H. Bouk, "Enhanced prediction of stock markets using a novel deep learning model PLSTM-TAL in urbanized smart cities," *Heliyon*, vol. 10, no. 6, Mar. 2024, Art. no. e27747.

[46] R. Zhang, C. Zhao, and G. Lin, "Interpretable image-based deep learning for price trend prediction in ETF markets," *Eur. J. Finance*, vol. 31, pp. 1–29, Nov. 2023, doi: 10.1080/1351847x.2023.2275567.

[47] C.-B. Ju and A.-P. Chen, "Identifying financial market trend reversal behavior with structures of price activities based on deep learning methods," *IEEE Access*, vol. 10, pp. 12853–12865, 2022, doi: 10.1109/ACCESS.2022.3146371.

[48] G. Wojarnik, "The potential of convolutional neural networks for the analysis of stock charts," *Proc. Comput. Sci.*, vol. 225, pp. 941–950, Jan. 2023, doi: 10.1016/j.procs.2023.10.081.

[49] M.-C. Hung, A.-P. Chen, and W.-T. Yu, "AI-driven intraday trading: Applying machine learning and market activity for enhanced decision support in financial markets," *IEEE Access*, vol. 12, pp. 12953–12962, 2024, doi: 10.1109/ACCESS.2024.3355446.

[50] N. Gârleanu and L. H. Pedersen, "Dynamic trading with predictable returns and transaction costs," *J. Finance*, vol. 68, no. 6, pp. 2309–2340, Dec. 2013.

[51] D. Araci, "FinBERT: Financial sentiment analysis with pre-trained language models," 2019, *arXiv:1908.10063*.

[52] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 12, May 2021, pp. 11106–11115.

IBANGA KPEREOBONG FRIDAY (Member, IEEE) received the B.Sc. degree in computer science (technology) from Babcock University, Ilishan-Remo, Nigeria, in 2017, and the M.Sc. degree from PDM University, Haryana, India, in 2021. He is currently pursuing the Ph.D. degree with the Department of Science and Engineering, Siksha 'O' Anusandhan (Deemed to be) University, Odisha, India. He has published seven papers in different conferences and journals. His research interests include stock market trend prediction, deep learning, and time-series analysis.

SARADA PRASANNA PATI is currently a Professor with the Department of Computer Science and Engineering, SOA University. He has 22 years of experience in teaching both undergraduate and postgraduate students of computer science and engineering. He has authored two textbooks and has published more than 20 research articles in various reputed journals and conference proceedings. His area of research interests include machine learning, deep learning, soft computing, medical image processing, and trend analysis.

DEBAHUTI MISHRA (Senior Member, IEEE) received the M.Tech. degree in computer science and engineering from KIIT Deemed to be University, Bhubaneswar, India, in 2006, and the Ph.D. degree from Siksha 'O' Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India, in 2011. She is currently a Professor and the Head of the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be) University. She has authored eight books and over 260 research papers. Her research interests include data mining, financial market prediction, and image processing.