LSTM-based Stock Prediction and Investment Model

Name	Student Number	Class	Division of Labor
陈思瑜	202464870046	AI 1	Early Attempts (Random Forest, XGBoost), Model Training, Strategy Evaluation, Visualizaion
柯咏曦	202430330863	AI 1	Early Attempts (XGBoost), Model Correction, Result Analysis, Code Management, Report Writing, Web Design
赵馨柠	202464831422	AI 1	Early Attempts (Random Forest), Data Preprocessing, Model Training, Strategy Evaluation, Report Repair
罗淇玥	202430420889	AI 1	Early Attempts (LightGBM, Random Forest), Model Training, Feature Analysis
钱雅萱	202430410679	AI 2	Early Attempts (LightGBM, Random Forest, LSTM), Model Training, Project Integration, Report Repair

Table 1 Team Member Information and Division of Labor

1 Problem Analysis

1.1 Problem restatement

Given an initial capital of 100,000 RMB and an investment period from January 1, 2024, to April 24, 2025, this project aims to construct an intelligent investment system integrating machine learning and quantitative trading strategies based on the historical daily data of Tencent Holdings. The model is trained using historical data from 2018 to 2023.

(1) Perform data cleaning, normalization, and missing value imputation on raw daily data. On this basis, extract technical indicators, time-series features, volatility features, and potential macro or sentiment proxies to construct a multi-dimensional feature set with predictive power.

(2) Design and implement a machine learning model suitable for stock price prediction, utilizing 2018–2023 data as the training set to forecast price trends for 2024–2025.

(3) Formulate explicit buy/sell signal rules based on model forecasts to build a complete trading strategy. The goal is to achieve alpha (excess returns) while tracking or outperforming the benchmark.

(4) Strictly control the maximum drawdown (MDD) of the portfolio while pursuing returns.

(5) Visualize the final investment process and results.

1.2 Analysis of the Problem

The core of this problem lies in data cleaning and feature selection of raw financial time-series, followed by the construction, training, and optimization of an appropriate predictive model. Simultaneously, an effective trading strategy must be designed to simulate the decision-making process in a real-world investment environment.

This task is a multi-objective optimized supervised learning problem. The modeling process requires a choice between regression and classification models. Traditional methods often employ tree-based models, such as Random Forest and XGBoost, after preprocessing and feature engineering, due to their stability and generalization capabilities on structured data. However, stock data exhibits significant temporal dependencies and sequential characteristics, which limit traditional models' ability to capture long-term patterns. Given these traits, this paper chooses Long Short-Term Memory (LSTM) networks to model stock time-series, fully mining the temporal correlations within historical prices and technical indicators.

Furthermore, the dataset is partitioned into training, validation, and test sets. Evaluation metrics such as R^2, RMSE, and Directional Accuracy are introduced on the validation set to calculate a weighted score for different models, selecting the highest performer for prediction. This selection process effectively mitigates overfitting and enhances generalization on unseen data. Regarding the trading strategy, the key challenge is converting predictions into executable decisions. Since relying solely on model outputs or traditional indicators has limitations, we propose a hybrid strategy that integrates both. Additionally, ablation studies are conducted to evaluate the specific contribution of each component to profitability and risk control.

1.3 Model Assumptions

To simplify the trading logic and focus on model performance, the following assumptions are made:

（1）Transaction costs, including commissions and stamp duties, are ignored.

（2） All trades are executed at the daily closing price; sell orders can only be executed at the close if a position is held.

（3）The market has sufficient liquidity, allowing immediate execution at the closing price without considering slippage or market impact costs.

2 Model Construction

2.1 Data Preprocessing

To ensure accuracy and continuity, raw data underwent rigorous cleaning: converting dates to standard time-series format, sorting chronologically, and handling outliers and missing values. The cleaned data was saved in CSV format.

2.2 Feature Engineering

2.2.1 Factor Mining and Feature Extraction

By integrating financial domain knowledge with quantitative investment logic, we extracted a feature system comprising 10 core features across three dimensions: price patterns, trend tracking, and momentum oscillations.The first dimension is Raw Price Data, which includes Open, High, Low, and Close prices, along with Trading Volume. These represent the most fundamental temporal information of market transactions, where volume serves as a proxy for the intensity of market sentiment. The second dimension is Trend Indicators, where we calculated the 5-day Moving Average (MA5) and the 20-day Moving Average (MA20). These are utilized to characterize short-term market sentiment fluctuations and medium-to-long-term support and resistance levels, respectively. The final dimension covers Oscillation and Momentum Indicators. By constructing the MACD (Moving Average Convergence Divergence) system, we extracted the fast line (DIF), the slow line (DEA), and the momentum histogram (MACD_Hist). These features aim to capture the "acceleration" characteristics of price changes, assisting the model in identifying trend reversals and continuations.

2.2.2 Correlation Analysis

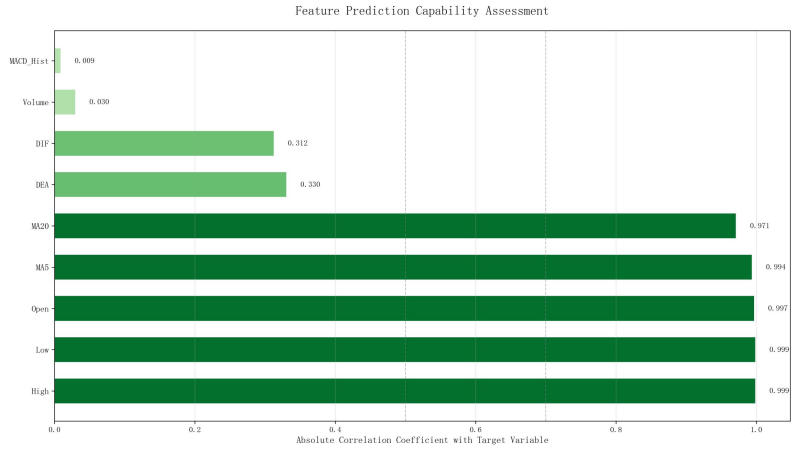

Figure 2-1 Correlation Analysis between Features and the Target Variable (close)

In the assessment of feature predictive potency, experimental results indicate that price-based features (High, Low, Open) and Moving Averages (MA5, MA20) exhibit extremely high predictive correlations with the target, with all correlation coefficients exceeding 0.97. Notably, the correlations for High and Low prices reach as high as 0.999.This suggests that short-term moving averages and historical price levels provide powerful linear guidance for the next trading day's closing price, forming the core foundation for the LSTM model to capture trend continuity.By contrast, the correlation coefficients for the momentum indicators DEA and DIF are 0.330 and 0.312, respectively. Although their linear correlation is relatively low, they provide kinetic transition information that is independent of absolute price levels.The correlation coefficients for Trading Volume and the MACD Histogram (MACD_Hist) are only 0.030 and 0.009, respectively. This implies that these indicators do not maintain a simple linear relationship with price; rather, their value is primarily reflected in the anticipation of non-linear trend reversals.

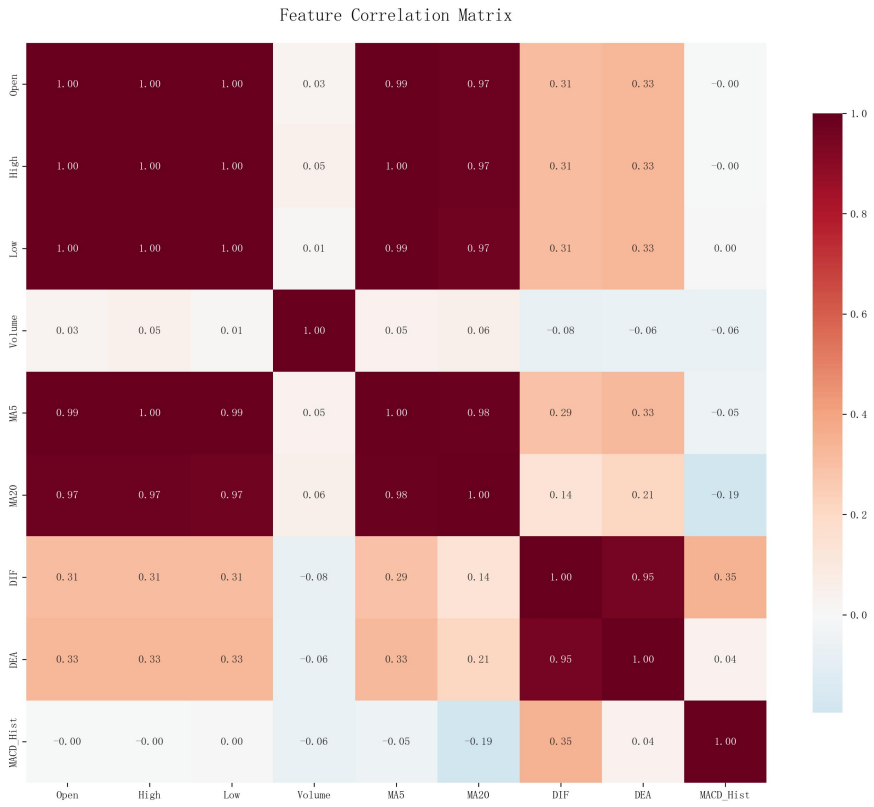

Figure 2-2 Feature Correlation Matrix

Regarding the analysis of feature redundancy, the correlation matrix reveals significant collinearity. For instance, the correlation coefficients between High, Low, and Open prices, as well as between MA5 and other price-based features, are all near 1, demonstrating a high degree of synchronicity.Despite the high level of redundancy among features, we opted not to simply remove them. Instead, we retained these indicators for their explicit financial

significance, allowing the model to automatically learn and extract effective information during the sequential modeling process.

In the model training phase, to eliminate the impact of different feature scales on the stability of deep learning training, we applied Min-Max Normalization to all input features, mapping them into a uniform range of [0, 1].Simultaneously, we constructed samples using a sliding time window with a length of 15 trading days based on the normalized feature sequences. This approach fully leverages the advantages of LSTM in modeling both long-term and short-term dependencies, enabling the extraction of predictive non-linear patterns from highly correlated temporal features.

2.3 Prediction Model

In the construction of the predictive engine, we evaluated multiple machine learning algorithms to address the high noise and non-linear characteristics inherent in stock price data. Although ensemble tree models, such as Random Forest, XGBoost, and LightGBM, demonstrate robust fitting capabilities and stability when processing structured data, their modeling mechanisms primarily rely on static non-linear combinations of features. Consequently, they struggle to effectively capture the continuously evolving dynamic characteristics typical of financial time series. In contrast, Long Short-Term Memory (LSTM) networks possess an inherent advantage in capturing market "memory effects" and temporal dependencies by modeling the impact of historical information on future trends. Therefore, we ultimately selected LSTM as the core predictive model.

During the model training and selection process, we introduced a rigorous validation-driven mechanism to ensure that the hyperparameter tuning process remained independent, practically feasible, and generalized. Specifically, data from July to December 2023 was partitioned into an independent validation set. This time interval was physically isolated during the training phase, ensuring it remained completely excluded from the parameter learning process, thereby effectively mitigating potential data leakage risks.Building upon this foundation, ten representative parameter combinations—varying in learning rates, dropout ratios, and the number of hidden LSTM units—were constructed and evaluated uniformly on the validation set. Model performance was measured using a weighted comprehensive scoring function, where R^2 and RMSE each accounted for 30%, and directional accuracy accounted for 40%. This approach balanced numerical fitting precision with trend judgment capability, allowing us to select the model that performed best on unseen data for subsequent investment decision-making.

2.4 Trading Strategy Construction

2.4.1 Indicator Selection

In the process of constructing the trading strategy, we integrated traditional statistical indicators from finance with the predictive outputs of deep learning models. This approach aims to introduce data-driven predictive power while maintaining model interpretability. The specific indicators selected and their respective roles are as follows:

- MA5 (5-Day Moving Average):

As a fundamental trend-following tool, the core principle of the MA5 is to reveal recent market average costs and directional preferences by smoothing short-term price fluctuations. When the closing price consistently remains above the MA5, it typically indicates that short-term buying pressure is dominant and the market is in an uptrend; conversely, it suggests strengthening selling pressure. We utilize the MA5 to identify the direction of short-term momentum.

- MACD (Moving Average Convergence Divergence):

The MACD consists of three components: the DIF, the DEA, and the MACD histogram. The DIF is the difference between the 12-day and 26-day Exponential Moving Averages (EMA), while the DEA is the 9-day exponential smoothing of the DIF. We primarily leverage the relative positioning of the DIF and DEA to assess momentum strength. When the DIF is higher than the DEA, it signifies strengthening bullish momentum, serving as a potential buy signal; when the DIF is lower than the DEA, it indicates dominant bearish momentum, constituting a sell or "wait-and-see" signal. This indicator not only reflects trend direction but also characterizes the acceleration of trend changes to some extent, offering leading significance for identifying periodic market reversals.

- LSTM Model Predicted Value:

The LSTM model performs regression forecasting of the next day's closing price based on historical price and technical indicator sequences. This study further utilizes the predicted percentage change relative to the current price as a critical basis for trading decisions. Because LSTM effectively captures non-linear relationships and long-term dependencies within financial time series, its predictions reflect potential short-term evolutionary trends, providing supplementary information for complex patterns that traditional technical indicators struggle to characterize.

2.4.2 Trading Strategy Design

In this study, the core scheme developed is the "LSTM + MA5 + MACD" hybrid strategy. The design philosophy behind this strategy is to construct a multi-dimensional

decision-making framework that integrates short-term trend assessment with a momentum confirmation mechanism.

The strategy relies on two classical technical indicators: first, whether the price sustains above the 5-day Moving Average (MA5) to identify short-term localized uptrends; and second, whether the DIF line in the MACD indicator is positioned above the DEA line, reflecting the dominance of bullish momentum. A buy signal is triggered whenever either of these conditions is met (Logical OR). In contrast, a sell signal—resulting in full liquidation—is executed only when the price falls below the MA5 and the DIF is simultaneously lower than the DEA (Logical AND).This architecture ensures the strategy remains highly responsive to market shifts, enabling early entry as a trend begins to emerge. By employing a "Logical OR" mechanism for entry, the system maintains a sensitive capture of potential opportunities. Simultaneously, the exit logic imposes more stringent confluence conditions, aiming to filter out "whipsaws" and ineffective trades in sideways markets. This enhances position stability and mitigates premature exits caused by minor volatility within an overall upward trend.

3 Solution and Result Analysis

3.1 Model Performance Evaluation and Selection

Following the feature construction and LSTM-based modeling, the objective and robust selection of the final predictive model for investment decision-making becomes a critical issue.

Since a single evaluation metric often fails to fully reflect the practical value of a model in financial scenarios, we treat the model selection process as a multi-criteria comprehensive evaluation problem.

We evaluate model performance on the validation set across three dimensions, covering both predictive accuracy and trading utility:

- R^2 (Coefficient of Determination)

  Used to measure the overall fitting capability of the model regarding real price fluctuations, reflecting the explanatory power of the regression.

- RMSE (Root Mean Square Error)

  Describes the absolute error level between predicted and actual values, measuring the stability and precision of the model.

- Directional Accuracy

  Calculates the proportion of correct predictions regarding the direction of future price

movements. This metric directly impacts the effectiveness of buy/sell decisions in a trading strategy and holds higher practical significance in financial applications.

Based on these metrics, we constructed a weighted comprehensive scoring function to unify and integrate the evaluation, where R^2 and RMSE each account for 30%, and Directional Accuracy accounts for 40%. By assigning a higher weight to Directional Accuracy, we ensure the model selection aligns more closely with the real-world trading requirement that "directional judgment prioritizes amplitude fitting."

During the training process, we designed multiple parameter combinations for different LSTM architectures—including hyperparameters such as the number of LSTM units, dropout rates, and learning rates—and evaluated each on the validation set.Ultimately, the model with the highest comprehensive score was selected as the optimal predictive model and applied to price forecasting and trading strategy backtesting on the test set.

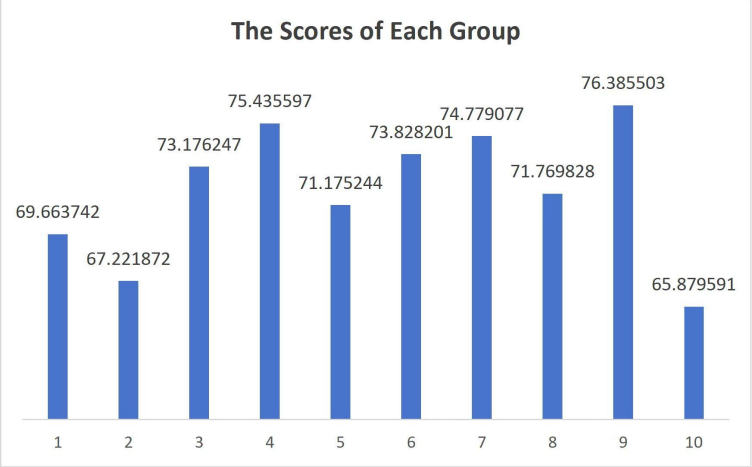

Figure 3-1 Bar chart of scores for each model

Model Selection: Parameter set 9 (Units=128, Dropout=0.4, LR=0.0005) achieved the highest score. On the test set, the model followed price trends closely with R^2=0.97 and RMSE=12.48.

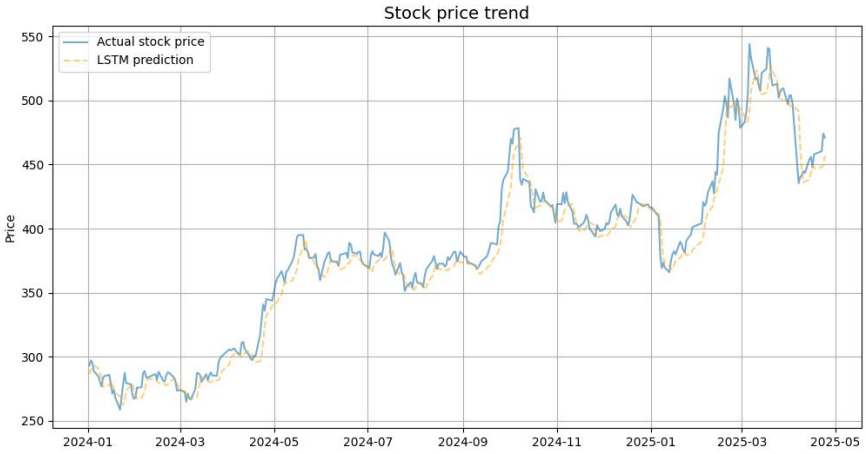

Figure 3-2 Comparison between Actual Stock Prices and LSTM Predicted Results ($R^2 = 0.97$, RMSE = 12.48)

We apply the model with the highest composite score to the test set. As shown in Figure 3-2, the model is able to closely track the true price movements in terms of both the overall trend and short- to medium-term fluctuations. Although a certain degree of lag can be observed, the overall fitting performance remains satisfactory. The model achieves a coefficient of determination of $R^2 = 0.97$ and a root mean squared error of RMSE = 12.48. These results indicate that the proposed model exhibits high accuracy and stability in the price prediction task, thereby providing a reliable informational basis for the subsequent construction of trading strategies based on its forecasts.

3.2 Trading Strategy Evaluation

To quantitatively assess the contribution of each component within the proposed hybrid strategy and to verify the necessity of incorporating LSTM-based predictive signals, a series of ablation experiments were designed. We first consider a buy-and-hold strategy as the baseline benchmark, in which the asset is purchased on the first day of the test period and held until the end without any intermediate rebalancing. This strategy effectively reflects the natural price appreciation of the underlying asset over the test interval and serves as a lower-bound reference for evaluating whether active trading strategies truly generate excess returns.

Building upon this baseline, we construct a moving-average-based strategy without LSTM signals (MA5 + MACD), which relies solely on price positions relative to the moving average and momentum comparisons between DIF and DEA. By comparing the performance of this strategy with that of the core hybrid strategy, we can clearly observe the incremental value introduced by LSTM-based predictions on top of traditional technical analysis.

To further investigate the predictive behavior of the LSTM model itself, several

comparative experiments are conducted. The pure LSTM regression strategy completely removes technical indicators and executes trades based only on fixed thresholds of +1% and −1.5%, aiming to evaluate the raw profitability of model outputs in the absence of any filtering mechanism. The pure directional strategy simplifies the logic even further by reducing the regression model to a binary classifier, focusing solely on the predicted direction of price movement rather than its magnitude, thereby testing the robustness of the strategy when signal strength information is discarded. In addition, the weak LSTM strategy increases the trading thresholds to ±3% to examine the model's accuracy under extreme, high-confidence signals and to verify whether stronger signals indeed correspond to higher predictive reliability.

Finally, to explore how different logical combinations of strategy components affect overall stability, two additional variants are constructed. The LSTM + MA5 strategy removes the MACD momentum confirmation module to assess performance when only short-term trend information is combined with LSTM predictions. The strict AND strategy requires simultaneous agreement among MA5, MACD, and LSTM signals before executing any trade. Through this multi-dimensional ablation framework, we not only validate the superiority of the proposed core strategy but also gain deeper insights into the specific contributions of each logical module to overall investment returns and risk control.

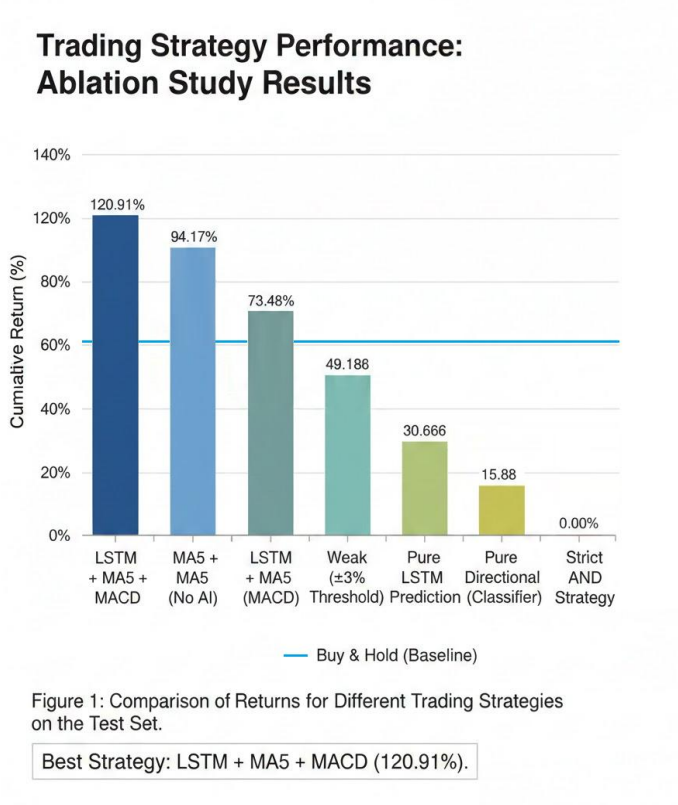

Figure 3-3 Comparison Chart of Returns for Different Trading Strategies on the Test Set

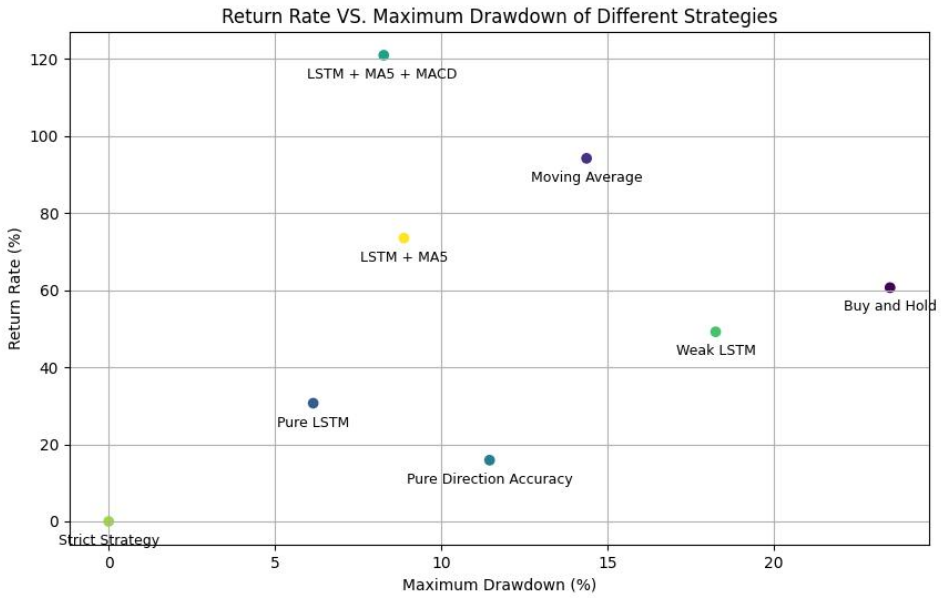

Figure 3-4: Maximum Drawdown vs. Returns of Different Trading Strategies

The experimental results show that the LSTM + MA5 + MACD hybrid strategy ranks first with a cumulative return of 120.91%, significantly outperforming the buy-and-hold benchmark return of 60.63%, thereby validating the superiority of the "forward-looking prediction + trend–momentum confirmation" framework. The solid performance of the buy-and-hold strategy establishes the bullish market regime during the test period, while the 94.17% return achieved by the moving-average strategy further demonstrates the practical effectiveness of MA5 and MACD in identifying short-term trends. On this basis, the core strategy realizes an additional 26.74% incremental return, which can be attributed to the LSTM model's ability to anticipate price turning points in advance. This enables the strategy to establish positions before trends are fully confirmed, achieving a successful complementarity between LSTM-based forecasts and traditional technical indicators.

A comparison of the ablation experiment results reveals a pronounced gap between prediction accuracy and profitability. The pure LSTM regression strategy (30.66%) and the pure directional strategy (15.88%), lacking effective filtering mechanisms, are overly sensitive to minor price fluctuations, resulting in frequent invalid trades and performance far inferior to the buy-and-hold benchmark. This indicates that relying solely on raw LSTM outputs makes the strategy highly vulnerable to noise. Even the weak LSTM strategy (49.18%), which attempts to capture high-confidence signals by raising trading thresholds, misses a large number of early trend opportunities due to excessive conservatism. Meanwhile, the strict AND strategy (0.00%) fails to trigger trades altogether because of overly restrictive entry conditions. These contrasts clearly expose the limitations of single-source signals: they

are either overly sensitive or completely ineffective.

Overall, the superiority of the core strategy lies in its construction of a scientific multi-dimensional decision space. It leverages LSTM to capture nonlinear price dynamics, uses MA5 to establish the prevailing market regime, and employs MACD to confirm momentum expansion. This OR-based entry logic maintains sensitivity while introducing multiple layers of fault tolerance, whereas the AND-based exit logic effectively filters false signals during sideways markets. By integrating the micro-level predictive capability of LSTM with the macro-level trend confirmation of technical indicators, the core strategy forms a complete logical loop of “prediction + confirmation + filtering”. This design effectively mitigates the tendency of artificial intelligence models to overfit noise in financial forecasting and achieves a crucial transition from numerical prediction to actionable trading value

3.3 Trading Performancec Analysis

Initial Capital	100,000.00
Final Capital	220,905.40
Total Return Rate	120.91%
Maximum Drawdown Rate	-8.27%
Number of Transactions	223

Table 2 Summary of Backtesting Results

Based on the backtesting results over the test set period, the proposed hybrid trading strategy combining trend indicators, technical indicators, and LSTM-based predictions achieves a notably strong overall performance. Using an initial capital of 100,000 as the benchmark, the strategy attains a final portfolio value of 220,905.40, corresponding to a total return of 120.91%. During the entire test interval, the strategy executes 223 trades (including both buy and sell operations), with a maximum drawdown of −8.27%. Overall, the strategy demonstrates a well-balanced trade-off between profitability, stability, and risk control, validating the effectiveness of integrating deep learning–based forecasts with traditional technical indicators.

Figure 3-5 Comparison of Strategy Net Value and Buy-and-Hold Benchmark

This figure presents a comparison between the net asset value curve of the proposed hybrid strategy and that of the buy-and-hold benchmark. It can be observed that the strategy's net value remains consistently and significantly above the benchmark throughout the test period. In particular, during phases of heightened market volatility, the strategy effectively mitigates drawdowns through active timing decisions, thereby achieving sustained excess returns over the benchmark.

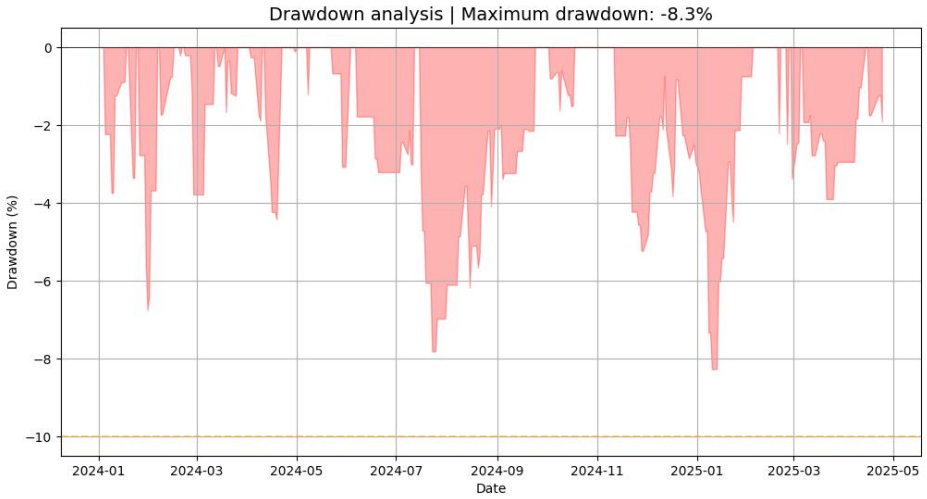

Figure 3-6 Strategy Drawdown Curve

This figure illustrates the risk exposure of the strategy during the testing period. The maximum drawdown is approximately −8.27%, and the duration of drawdowns is relatively limited, indicating that the strategy maintains effective risk control characteristics.

Figure 3-7 Illustration of Trading Timing

This figure overlays the buy and sell signals generated by the strategy on the actual stock price curve. The trading signals are predominantly concentrated around trend inflection points or acceleration phases, with buy and sell points appearing in a relatively dense yet structured manner, reflecting the strategy's sensitivity to trend changes and momentum shifts.

3.4 Web-based Visualization System

To present the model prediction results and trading strategy performance in a more intuitive and clear manner, the relevant outcomes are collectively displayed through a web-based visualization platform. The access details are provided as follows:

https://collinke05.github.io/Machine-Learning-Project-Stock-Investment-Analysis/%E5%A4%A7%E6%B5%AA%E6%B7%98%E9%87%91final/%E7%BB%93%E6%9E%9C%E5%8F%AF%E8%A7%86%E5%8C%96%E7%BD%91%E9%A1%B5/index.html

4 .Model Evaluation and Practical Deployment

The core strength of the proposed system lies in its rigorous validation-driven model selection mechanism. By conducting blind evaluations of ten parameter configurations on a physically isolated validation period, and selecting models based on a weighted composite score derived from R^2, RMSE, and directional accuracy, the system significantly enhances generalization performance on unseen test data while effectively mitigating the risk of overfitting. Experimental results demonstrate that the composite LSTM + MA5 + MACD strategy substantially outperforms the buy-and-hold benchmark, achieving a total return of

120.91% while maintaining the maximum drawdown at –8.27%. These findings confirm that the proposed “prediction–confirmation–filtering” framework achieves an excellent balance between return capture and risk control.

In terms of robustness, the system exhibits strong noise resistance and stability. By incorporating MA5-based trend filtering and MACD-based momentum confirmation, the strategy effectively suppresses minor prediction deviations produced by the LSTM model under range-bound or highly volatile market conditions. The ablation experiments further reveal a pronounced performance gap between the pure LSTM strategy and the hybrid strategy, highlighting that this multi-dimensional decision logic substantially improves the system’s fault tolerance and robustness.

Regarding scalability and applicability, the adopted “temporal features + general technical indicators” architecture demonstrates strong adaptability. While validated on highly liquid large-cap stocks such as Tencent Holdings, the framework can be readily extended to other industry sectors or index-level assets. In future work, the integration of sentiment-based features or more advanced attention-based neural architectures may further enhance the system’s ability to achieve stable excess returns across diverse and dynamically evolving market environments.