Advanced Data Processing and Predictive Modeling Techniques for Enhanced Quantitative Trading Performance

***Abstract*—In our work, we focus on advanced data preprocessing and predictive modeling to enhance quantitative trading efficiency. We detail essential steps like data reading, type conversion, handling of NaN values, standardization, and transformations specific to minute-level data, emphasizing consistency in data management. Our evaluation of various predictive models, such as Random Forest, GRU and LSTM, A2C and PPO, and AutoGluon, highlights their effectiveness in regression, time series analysis, and trading decision optimization. We discuss investment strategies that integrate daily and intraday data with technical indicators and risk management principles. Finally, our creation of a quantitative trading investment portfolio based on key Hang Seng Index stocks, analyzed using advanced tools, showcases the integration of comprehensive data processing and predictive modeling in enhancing trading decision-making and efficiency.**

I. Introduction

IN the evolving landscape of financial markets, the integration of advanced data processing and predictive modeling has become paramount in achieving success in quantitative trading. Our work addresses the growing need for sophisticated, data-driven strategies in the finance industry, leveraging state-of-the-art techniques in data preprocessing and machine learning. The objective is to optimize investment decisions, manage risk, and enhance profitability in a highly competitive environment.

The cornerstone of our approach lies in meticulous data preprocessing. Recognizing that the quality of input data significantly influences the performance of predictive models, we employ rigorous steps such as data type conversion, handling missing values, and applying various transformations. These steps ensure data integrity and prepare it for effective analysis and modeling.

Our exploration of predictive models like Random Forest, GRU and LSTM, A2C, PPO, and AutoGluon reflects the diverse range of tools available for quantitative analysis. Each model offers unique advantages in handling different aspects of financial data, from regression problems to time series forecasting and decision optimization in trading. By comparing and contrasting these models, we aim to identify the most effective techniques for various trading scenarios.

Furthermore, our work delves into the development of comprehensive investment strategies. These strategies are designed to capitalize on the strengths of predictive models while addressing the complexities of financial markets. We emphasize the use of both daily and intraday data, integrating technical indicators, and implementing robust risk management practices. This holistic approach ensures that our strategies are not only grounded in strong theoretical foundations but are also practical and adaptable to real-world trading conditions.

Finally, the creation of a quantitative trading investment portfolio, based on key stocks from the Hang Seng Index, demonstrates the application of our methodologies in a practical setting. This portfolio is analyzed using modern analytical tools, providing insights into its performance and the effectiveness of our strategies.

II. Data Preprocessing

A. Combined Data Preprocessing Steps

- **Reading Data:** Read data (daily or minute-level) from the specified file path, extracting column names and skipping the first line (header).
- **Data Type Conversion:** Convert all columns to numeric types, turning unconvertible values into NaN.
- **Handling NaN Values in Date Column:** Transform illegal date values into NaN.
- **Handling Missing Values:** Fill in NaN values using linear interpolation.
- **Data Standardization:** Standardize columns like opening price, highest price, etc., using `StandardScaler` to achieve zero mean and unit variance.
- **Log Transformation:** Apply log transformation to closing prices, ensuring all values are positive for handling skewed distributions.
- **Box-Cox Transformation:** Perform Box-Cox transformation on trading volume to improve normality or symmetry.
- **Additional for Minute-Level Data:** Compute a moving average (e.g., 5-minute moving average for closing prices) to smooth out short-term fluctuations in minute-level data.

B. Rationale Behind Data Processing

- **Data Type Conversion and Missing Value Handling:** Essential for ensuring data consistency and completeness, crucial for machine learning models.
- **Standardization:** Eliminates the impact of different scales, making model training more manageable.
- **Log and Box-Cox Transformations:** Transform skewed distributions to be more normally distributed, enhancing model performance.
- **Moving Average (for Minute-Level Data):** Effective in time series data for highlighting trends by smoothing out short-term fluctuations.

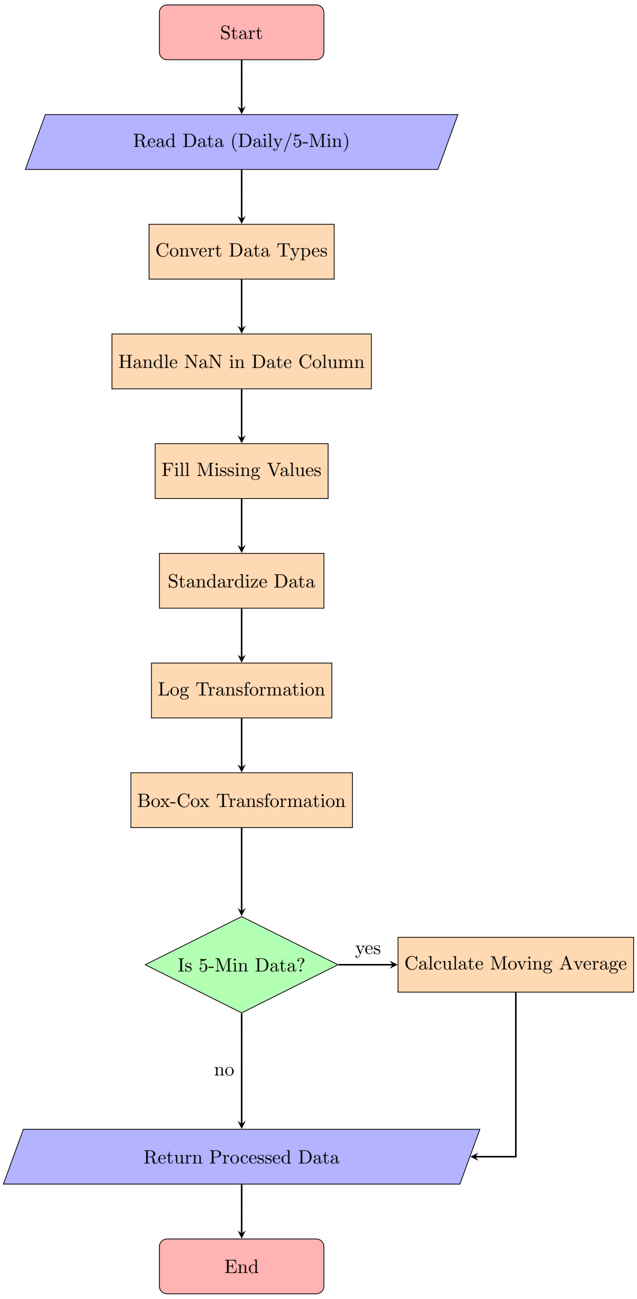

Fig. 1. Preprocessing Flowchart

III. Alogorithms & Methods

To predict the data, we attempted and compared multiple models. We used the data from November 17, 2020, to November 17, 2023, as the test set, and the data outside this range as the training and validation sets. The division of the training and validation sets was done using cross-validation methods. Below are the algorithms we used and their corresponding experimental results.

A. Random Forest

The Random Forest algorithm is a widely used ensemble learning method in machine learning, primarily known for improving prediction accuracy and controlling overfitting by building multiple decision trees and combining their outcomes. This algorithm employs bootstrap sampling to randomly select samples from the original dataset, providing different subsets of data for constructing each tree. During the tree-building process, it also randomly selects features to determine the splits at nodes, enhancing model diversity and reducing the risk of overfitting. In regression problems, it takes the average of the predictions from each tree. This approach not only increases the model's accuracy but also maintains a level of interpretability, helping in understanding the data by providing an assessment of feature importance.

The results of Random Forest algorithm are shown below:

TABLE I
Random Forest Performance Metrics

Metric	Value
MAPE (5min)	0.0459
R^2 (5min)	0.9839
MAPE (Daily)	0.3389
R^2 (Daily)	0.9880

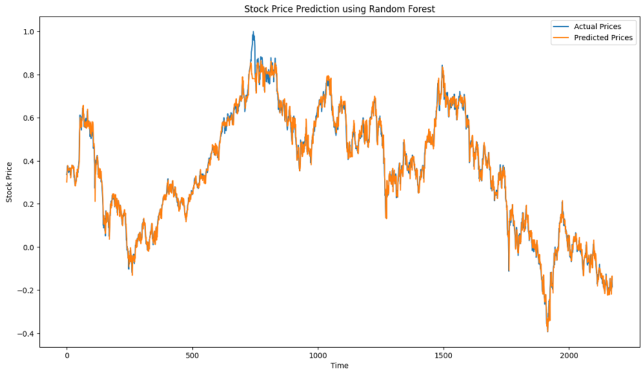

Fig. 2. Random Forest Prediction (Daily)

B. GRU & LSTM

GRU (Gated Recurrent Unit) and LSTM (Long Short-Term Memory) are both variants of recurrent neural networks (RNNs) used for processing sequential data. They effectively manage and convey information by introducing gating mechanisms, thus achieving strong performance in various time series tasks.

The results of GRU & LSTM algorithms are shown in Table II and Table III:

C. A2C & PPO

A2C and PPO are two popular reinforcement learning algorithms. A2C is utilized to optimize investment decisions by having an Actor generate trading strategies and a Critic estimate their value, making it suitable for quantitative trading. On the other hand, PPO specializes in policy optimization and offers enhanced training stability, making it valuable for

TABLE II
GRU PERFORMANCE METRICS

Metric	Value
MAPE (5min)	0.0024
R^2 (5min)	0.9996
RMSE (5min)	80.4224
MAPE (Daily)	0.01498
R^2 (Daily)	0.9839
RMSE (Daily)	467.0141

TABLE III
LSTM PERFORMANCE METRICS

Metric	Value
MAPE (5min)	0.0039
R^2 (5min)	0.9986
RMSE (5min)	150.6185
MAPE (Daily)	0.0034
R^2 (Daily)	0.9982
RMSE (Daily)	171.3058

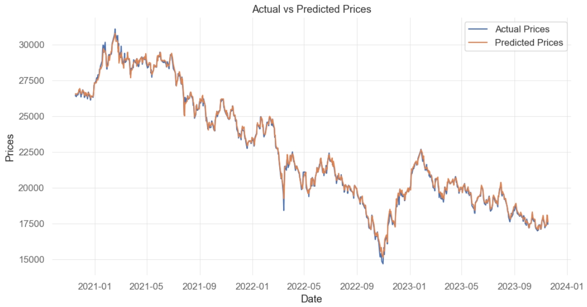

Fig. 3. LSTM Prediction (Daily)

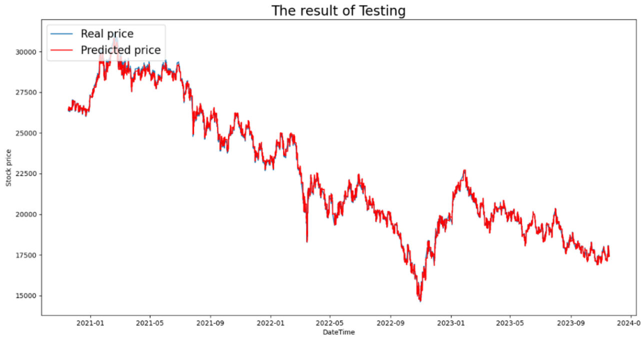

Fig. 4. GRU Prediction (5min)

improving trading strategies, managing investment portfolios, and enhancing risk control.

The results are shown in Fig 5.

D. AutoGluon

AutoGluon is an open-source automatic machine learning (AutoML) library developed by a team of Amazon scientists, aimed at simplifying the development and optimization of machine learning models. The core strength of AutoGluon lies in its ability to automate tedious processes such as data preprocessing, feature engineering, model selection, and hyperparameter tuning. It improves predictive performance by automatically evaluating and integrating the strengths of various models, including decision trees and deep learning models.

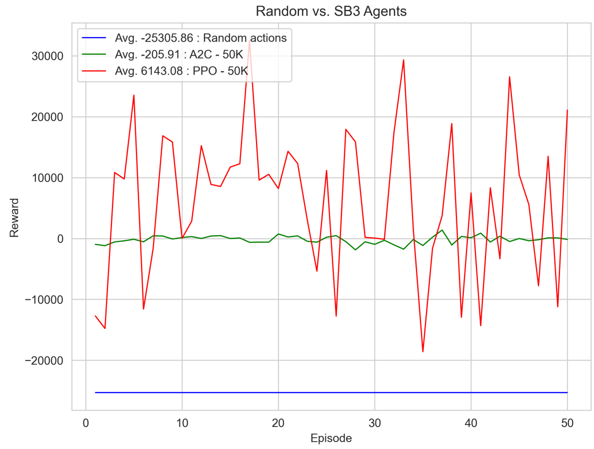

Fig. 5. A2C & PPO Agents (Daily)

However, due to the limit of coputational resources, we choosed the relatively simple models to predict. Here are some metrics of the trained model.

TABLE IV
AUTOGLUON PERFORMANCE METRICS

Metric	Value
MAPE (Daily)	0.0538
MSE (Daily)	0.0081

E. Model Comparison

TABLE V
MODEL COMPARISON

Model	MAPE	R^2
Random Forest (5min)	0.0459	0.9839
GRU (5min)	**0.0024**	**0.9996**
LSTM (5min)	0.0039	0.9986
Random Forest (Daily)	0.3389	0.9880
GRU (Daily)	0.0150	0.9839
LSTM (Daily)	**0.0034**	**0.9982**

IV. INVESTMENT STRATEGIES AND ASSESSMENT

A. Investment Strategies & Results

a) Trading Strategy Overview:

- **Data Input**: Utilize both daily and 5-minute interval data, focusing on actual values rather than model predictions.
- **Technical Indicator Calculation**: Emphasize the use of 20-period and 50-period Simple Moving Averages (SMAs) for signal generation.

b) Five-Minute Model: Trading Signals:

- **Buy Signal**: Issued when the current price is a certain percentage above the predicted closing price (set by buy_threshold) and the 20-period SMA is above the 50-period SMA.
- **Sell Signal**: Issued when the current price is below the predicted closing price by a set percentage (sell_threshold) and the 20-period SMA is below the 50-period SMA.

c) Trade Size and Risk Management:

- **Trade Size**: Maintain a fixed trade amount, adjustable via fixed_trade_amount.
- **Stop Loss and Take Profit**: Implement dynamic stop loss (stop_loss_pct) and take profit (stop_gain_pct) conditions to protect investments and lock in profits.

d) Daily Model: Trend Prediction:

- Retain the focus on forecasting daily average index values.
- Continue using predicted values for short-term and long-term moving average calculations.
- Maintain trend analysis based on SMA comparisons, but with enhanced focus on their role in trade signal generation.

e) Combining Models for Overall Strategy:

- **Long-Term Trend** (Daily Model): Continues to dictate the primary trading direction.
- **Intraday Signals** (5-Minute Model): Refined to incorporate specific conditions for buying and selling, based on price action and SMA relationships.
- **Execution Strategy**: Align intraday signals with the long-term trend, with a focus on dynamic risk management and performance evaluation, including the calculation of Net Asset Value (NAV) and total returns.

Using the HSI for trading and our trading strategy, the results obtained are as follows:

TABLE VI
AUTOGLUON PERFORMANCE METRICS

Metric	Value
Cumulative Return	0.8538
Max Drawdown	1.8813×10^{-6}

B. Building a Quantitative Trading Investment Portfolio

The portfolio includes key Hang Seng Index stocks: HSBC Holdings (10%), CNOOC Limited (50%), China Mobile (20%), and Nongfu Spring (20%). This diverse mix across finance, energy, telecom, and consumer sectors aims to capture the index's market dynamics.

Using pandas, we process historical stock data to compute daily returns for each stock. These returns are then weighted according to their portfolio allocation.Quantstats is used for in-depth performance analysis, providing metrics like Cumulative Return, CAGR, Sharpe Ratio, and Maximum Drawdown. These indicators help assess the portfolio's risk-adjusted returns and market resilience.

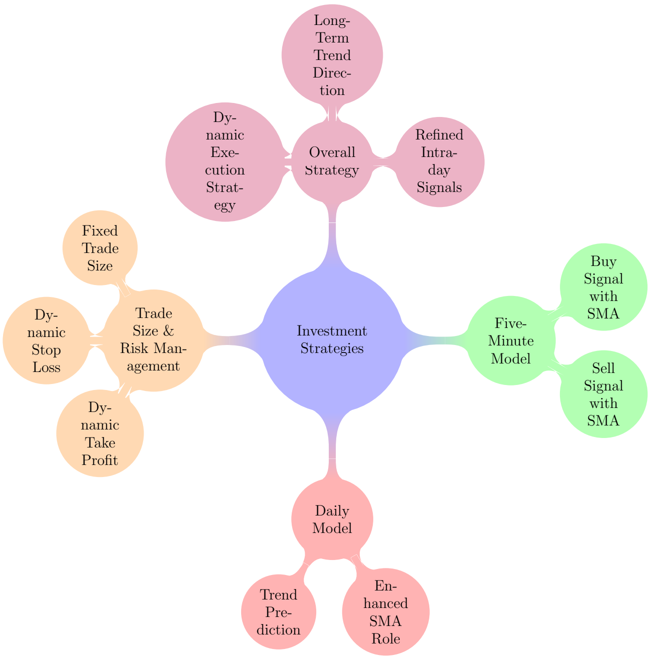

Fig. 6. Investment Strategies

The results of our investment portfolio is shown in Table VII.

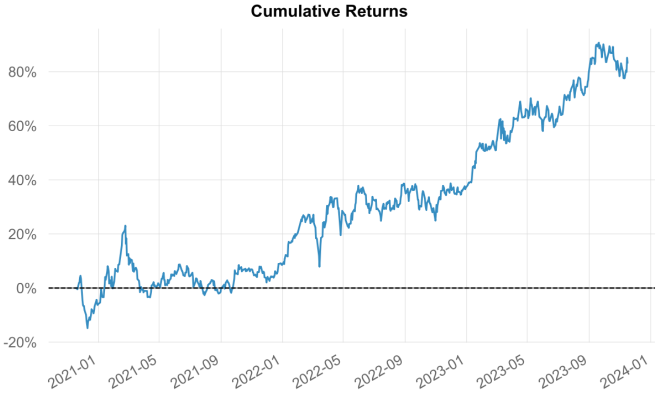

Fig. 7. Cumulative Returns of Investment Portfolio

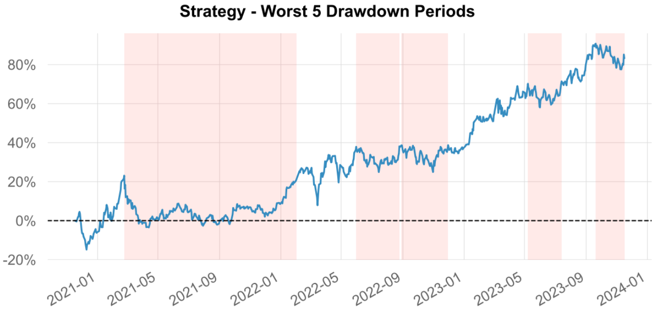

Fig. 8. Worst 5 Drawdown Periods of Investment Portfolio

TABLE VII
INVESTMENT PORTFOLIO PERFORMANCE METRICS

Metric	**Value**
Start Period	2020-11-17
End Period	2023-11-17
Risk-Free Rate	0.0%
Time in Market	100.0%
Cumulative Return	**80.05%**
CAGR	14.51%
Sharpe	**1.17**
Prob. Sharpe Ratio	97.71%
Sortino	1.74
Sortino/$\sqrt{2}$	1.23
Omega	1.22
Max Drawdown	**-15.52%**
Longest DD Days	331
Gain/Pain Ratio	0.22
Gain/Pain (1M)	1.37
Payoff Ratio	1.1
Profit Factor	1.22
Common Sense Ratio	1.43
CPC Index	0.7
Tail Ratio	1.17
Outlier Win Ratio	3.47
Outlier Loss Ratio	3.76
MTD	3.03%
3M	2.38%
6M	7.23%
YTD	33.6%
1Y	37.9%
3Y (ann.)	16.9%
5Y (ann.)	14.51%
10Y (ann.)	14.51%
All-time (ann.)	14.51%
Avg. Drawdown	-3.26%
Avg. Drawdown Days	30
Recovery Factor	4.12
Ulcer Index	0.07
Serenity Index	0.93

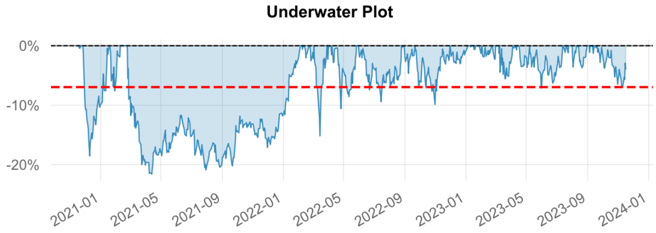

Fig. 9. Underwater Plot of Investment Portfolio

V. Conclusion

In summary, our study presents a pragmatic approach to integrating advanced data processing and predictive modeling in quantitative trading. Our efforts in data preprocessing established a reliable foundation for model application, emphasizing the importance of data quality in financial analysis.

The exploration of various models like Random Forest, GRU and LSTM, A2C, PPO, and AutoGluon offered insights into their respective strengths and limitations in trading contexts. This understanding is crucial for financial practitioners who must choose models that best fit their specific data and strategy needs.

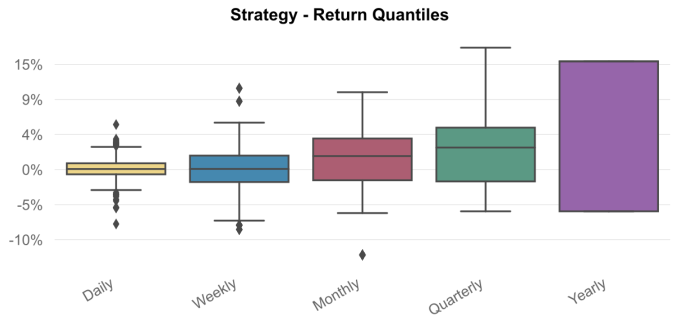

Fig. 10. Return Quantiles of Investment Portfolio

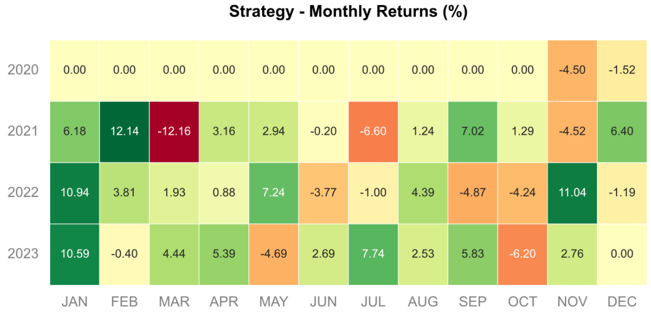

Fig. 11. Monthly Returns of Investment Portfolio

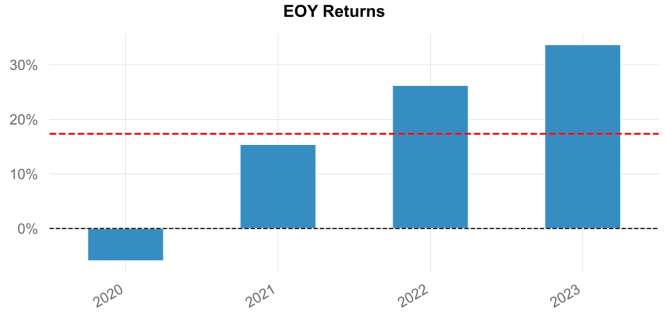

Fig. 12. EOY Returns of Investment Portfolio

The investment strategies we developed, utilizing both daily and intraday data, reflect a balanced approach to market analysis and risk management. These strategies, while theoretical in nature, provide a framework for practical application in dynamic market conditions.

Our analysis of a Hang Seng Index-based investment portfolio served as a case study for applying these methods in a real-world scenario. The performance of this portfolio, though limited in scope, suggests the potential effectiveness of our integrated approach.

This work contributes to the field of quantitative finance, offering a perspective on how data-driven techniques can be used in trading strategies. It underscores the ongoing need for adaptability and rigorous analysis in the face of ever-evolving financial markets.

Contribution

Runze Fang: A2C & PPO, Investment Portfolio and Strategy, Report Writing

Chengkai Wang: Preprocessing, LSTM & GRU, Investment Strategy

Weitai Sun: LSTM & GRU, Model Evaluating

Haichuan Wei: AutoGluon, Model Evaluating

Hongyu Long: Random Forest, Model Evaluating

Acknowledgements

I would like to express my sincere gratitude to Professor Liu and Professor Xu for their guidance and teaching in Machine Learning. Their advice has been greatly appreciated.