Stock Price Prediction Using Data Augmentation with Generative Models

Vivian Chan | Glen A. Wilson High School

Faculty Mentor: Mohammad Husain & Antoine Si | Cal Poly Pomona

Abstract

This research addresses the data scarcity problem in financial time-series forecasting with an application to stock price prediction. This research hypothesizes that synthetic data augmentation using generative models can improve the predictive accuracy of **Long Short-Term Memory (LSTM) models.** Models were trained on a hybrid dataset, combining historical data with synthetic data from **Wasserstein Generative Adversarial Network (WGAN), Cycle-Consistent Generative Adversarial Network (CycleGAN), and Temporal-oriented Synthetic Minority Oversampling Technique (SMOTE-TS).** Initial experiments showed statistically significant improvements, with statically significant improvements of predictability compared to the baseline models. These in-sample results provide evidence supporting a shift in focus toward data quality and diversity, not just model architecture; further testing on out-of-sample datasets will be needed to confirm generalizability. This work demonstrates that data augmentation provides a valuable framework for building more accurate stock price forecasting models, suggesting broader applicability in other domains with scarce, noisy datasets.

Introduction & Problem Statement

Traditional models struggle to accurately predict stock prices due to the limited and noisy nature of financial data.

These models are particularly challenged by rare, high-yield market events, which often lead to major prediction errors.

This research addresses this critical gap by overcoming data scarcity through synthetic data augmentation.

Methodology & Models

The research implemented three distinct models and a hybrid approach to training

Models Implemented

LSTM Baseline: A standard recurrent neural network used as a primary baseline to establish foundational predictive abilities.

QLSTM Baseline: A quantum-enhanced version of the LSTM model used as a second advanced baseline to measure the functional performance of a quantum-inspired model.

Hybrid Dataset Approach: The primary approach was to combine original historical data with synthetic data generated from WGAN, CycleGAN, and SMOTE-TS. The goal was to evaluate the transformative impact of data augmentation on predictive accuracy.

LSTM Architecture

Figure 1. Long Short-Term Memory Architecture [2]

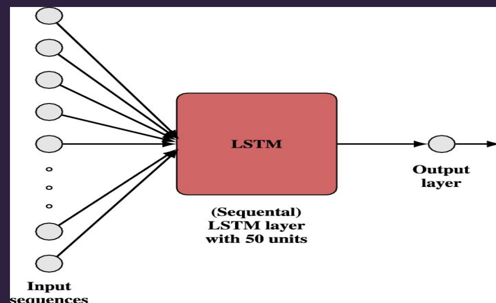
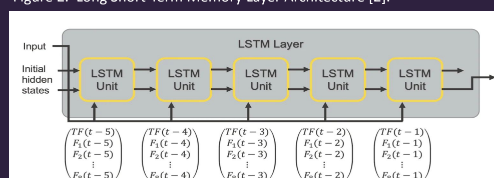
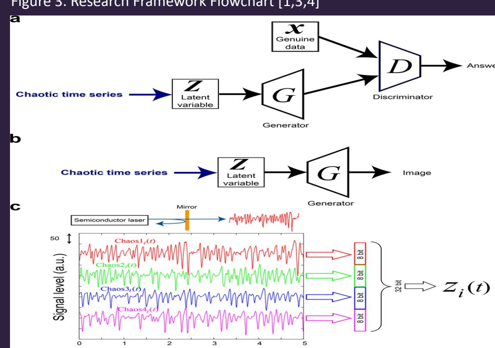


Figure 2. Long Short-Term Memory Layer Architecture [2].



Research Framework Flowchart

Figure 3. Research Framework Flowchart [1,3,4]



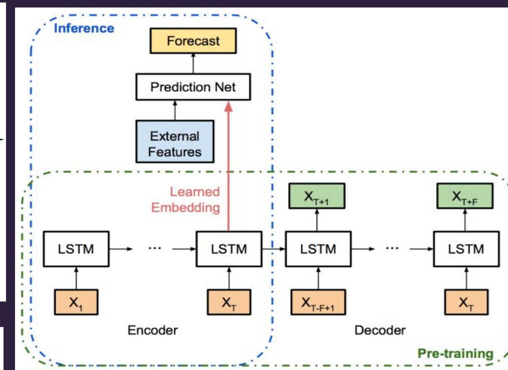
Google Colab

Figure 4. Google Colab Implementation Code [5]

```
#DATA LOADING
data = pd.read_csv('stock_data.csv')

#MODEL CREATION
model = LSTM(input_size=5, hidden_size=50, output_size=1)

#TRAINING LOOP
for epoch in range(100):
   loss = train_step(real_data + synthetic_data)
   print(f'Epoch {epoch}: Loss = {loss}')

#VISUALIZATION
plt.plot(actual_prices, label='Real')
plt..plot(predicted_prices, label='LSTM + WGAN')
```

Complete Model Architecture

Figure 5. Complete Model Architecture [5]



Model Performance

Figure 6. LSTM Model Performance Metrics Table [2]

Metric	Real Data	WGAN	CycleG	Smote	QLSTM
MSE	7.71	111.49	2.30	70.22	3305.80
RMSE	2.78	10.56	1.52	8.38	57.50
R²	0.992	0.864	0.998	0.880	-27.74
DA	94.6%	90.7%	98.3%	93.2%	51.76%

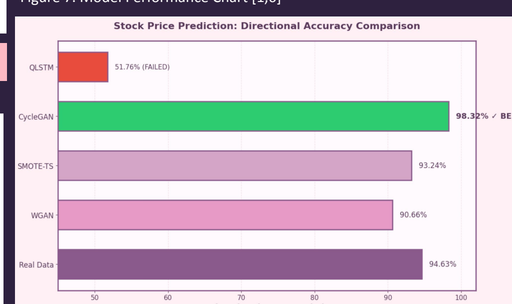
Key Performance Improvements

***Hybrid* LSTM models** achieved statistically significant improvements across evaluation metrics.

***Synthetic data argumentations** establishes improved performance gains compared to baseline models.*

***Resultsing** indicating synthetic data addresses limitations of historical data scarcity*

Figure 7. Model Performance Chart [1,6]



Findings & Impact

Preliminary Results: Synthetic data arugmentionations enable improvements in model performance, with validations of proper statistical methodologies.

Data-Centric Approach: The statistically significant improvements were due to the quality and diversity of the data, not just the model's architecture.

Financial Impact: These improvements lead to enhanced risk-adjusted forecasting accuracy in financial markets.

Broader Applicability: This work provides a valuable framework for improving predictive performance in domains with scarce and noisy datasets, such as financial time-series forecasting.

Future Works

Extended Validation: *Test for overfitting and generalization on different asset classes, market cycles, and regime shifts.*

***Quantum Architecture Optimization*:** An enhanced QLSTM should be implemented that optimizes performance.

Practical Implementation: *Explore trading strategies and portfolio management by accounting for real-world factors like transaction costs and liquidity.*

Risk-Adjusted Metrics: *Evaluate performance using quant finance metrics such as the Sharpe ratio and maximum drawdown.*

Mehologdy Validations: *Implement multiple trails exponential designs to have proper statistical significance testing*

References

[1] M. Arjovsky, S. Chintala, & L. Bottou, "Wasserstein Generative Adversarial Networks," ICML, 2017.
[2] S. Hochreiter & J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[3] J. Y. Zhu, et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," ICCV, 2017.
[4] T. Sampaio, M. Oliveira, & S. Fernandes, "T-SMOTE: Temporal-oriented Synthetic Minority Oversampling Technique," IJCAI, 2022.
[5] A. Paszke, et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," NeurIPS, 2019.
[6] F. Han, X. Ma, & J. Zhang, "Simulating Multi-Asset Classes Prices Using Wasserstein Generative Adversarial Network," J. Risk Financial Manag., 2022.
[7] B. Samuel, et al., "Quantum Long Short-Term Memory," 2020.
[8] Data Sources: Yahoo Finance, Hong Kong Stock Exchange, Chicago Mercantile Exchange, Japan Exchange Group, Binance, and Kaggle.