Project Overview

This research addresses the critical challenge of data scarcity in financial time-series forecasting through innovative synthetic data augmentation techniques. By leveraging advanced generative models (WGAN, CycleGAN, and SMOTE-TS) combined with LSTM neural networks, this project demonstrates statistically significant improvements in stock price prediction accuracy.

Key Results

Directional Accuracy: Improved from 44.31% to 99.58%
R^2 Score: Enhanced from 0.8823 to 0.9996
MSE Reduction: Decreased from 0.0035 to 0.0001
RMSE: Reduced from 0.0589 to 0.0082

Research Impact

Data-Centric Approach: Demonstrates that data quality and diversity are as crucial as model architecture
Financial Applications: Provides enhanced risk-adjusted forecasting accuracy for financial markets
Broader Applicability: Framework applicable to other domains with scarce, noisy datasets

Author Information

Vivian Chan
Glen A. Wilson High School
Faculty Mentors: Mohammad Husain & Antoine Si
Cal Poly Pomona

Repository Structure

```
stock-price-prediction/
│
├── README.md                          # Project documentation
├── notebooks/                         # Jupyter notebooks
│   └── PHASE_1__1___2_.ipynb         # Main implementation notebook
├── poster/                            # Research presentation
│   └── Chan_Vivian__10_.pdf          # Research poster
├── data/                              # Dataset files
│   └── (Historical stock data)
└── results/                           # Model outputs and visualizations
    └── (Performance metrics and plots)
```

Methodology

Models Implemented

LSTM Baseline: Standard recurrent neural network for foundational predictive capabilities
QLSTM Baseline: Quantum-enhanced LSTM model for advanced performance benchmarking
Hybrid Models: Combined historical data with synthetic data from:
Wasserstein GAN (WGAN)
Cycle-Consistent GAN (CycleGAN)
Temporal-oriented SMOTE (SMOTE-TS)

Technical Specifications

Framework: PyTorch
Environment: Google Colab with GPU acceleration
Training Parameters:

Batch size: 64
Learning rate: 0.001
Optimizer: Adam
Epochs: 100 with early stopping
Trials: 5 independent runs

Dataset

Primary Data: S&P 500 and Apple Inc. (AAPL) stock prices
Time Period: January 1, 2020 - January 1, 2023
Source: Yahoo Finance
Augmentation: Synthetic data generated using GANs and SMOTE-TS

Getting Started

Prerequisites

```
pythonpip install pandas numpy matplotlib
pip install torch torchvision
pip install yfinance
pip install scikit-learn
```

Running the Code

Clone this repository:

```
bashgit clone https://github.com/vnnviv/stock-price-prediction.git
cd stock-price-prediction
```

Open the Jupyter notebook:

```
bashjupyter notebook notebooks/PHASE_1__1___2_.ipynb
```

Run all cells to reproduce the results

Key Findings

Synthetic Data Augmentation Success

Hybrid models with synthetic data significantly outperformed baseline models
WGAN and CycleGAN demonstrated the most effective synthetic data generation
Addressed limitations of historical data scarcity, especially for rare high-yield events

Model Architecture Insights

Quantum-enhanced LSTM (QLSTM) showed minimal improvement without synthetic data
Data quality proved more impactful than architectural complexity
Hybrid approach validated across multiple independent trials

Future Work

Extended Validation: Test on different asset classes and market cycles
Quantum Architecture Optimization: Enhance QLSTM implementation
Risk-Adjusted Metrics: Incorporate Sharpe ratio and maximum drawdown
Practical Implementation: Develop trading strategies accounting for transaction costs
Methodology Validation: Implement multiple trial designs for statistical significance

References

M. Arjovsky, S. Chintala, & L. Bottou, "Wasserstein Generative Adversarial Networks," ICML, 2017

S. Hochreiter & J. Schmidhuber, "Long Short-Term Memory," Neural Computation, 1997

J. Y. Zhu, et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," ICCV, 2017

T. Sampaio, M. Oliveira, & S. Fernandes, "T-SMOTE: Temporal-oriented Synthetic Minority Oversampling," IJCAI, 2022

Yahoo Finance, Hong Kong Stock Exchange, Chicago Mercantile Exchange, Japan Exchange Group

Visualizations

The project includes comprehensive visualizations of:

Model performance comparisons

Training/validation loss curves

Directional accuracy improvements

Synthetic vs. real data distributions

Achievements

Statistically significant performance improvements across all metrics

Novel application of GANs to financial time-series augmentation

Validated framework for addressing data scarcity in volatile markets

----

This research was conducted as part of a high school research program under the mentorship of Cal Poly Pomona faculty.