Capstone Project Report

Submitted by: Aaryav Sharma (as18794), Haokang Mei (hm3235), Jiahao Liu (jl14869), Mansoor Alkaabi (maa1145), Shuqing Qi (sq2172), Xu Qi (xq665), Yang Xu (yx3267)

December 8, 2024

1 Introduction

The increasing complexity and volume of financial market transactions have made the identification of anomalous stock price movements both more crucial and more challenging. Market anomalies can arise from various sources, including market manipulation, technical glitches, or legitimate news events, making their detection and classification a significant challenge for market participants and regulators alike. While traditional statistical approaches have long been employed to identify such anomalies, the advent of machine learning techniques has opened new avenues for more sophisticated and accurate detection methods.

This paper presents a novel approach to detecting stock price anomalies by combining three powerful machine learning techniques: Variational Autoencoders (VAE), and Isolation Forest. Our methodology focuses specifically on identifying point outliers in daily stock returns, utilizing data from Russell 3000 constituent stocks over a ten-year period from 2014 to 2023. By employing a regression-based approach to isolate firm-specific returns from broader market movements, we create a more refined dataset for anomaly detection that minimizes the impact of market-wide trends.

The significance of our research lies in its innovative integration of complementary methodologies. The VAE, with its probabilistic framework, excels at learning compact representations of normal market behavior and identifying subtle deviations from these patterns. The Isolation Forest algorithm complements these approaches by efficiently identifying outliers in high-dimensional spaces. This tripartite combination allows for more robust detection of anomalies compared to traditional single-model approaches, particularly in the context of high-frequency, high-volume financial data where conceptual drift is common.

Our paper contributes to the existing literature by presenting a comprehensive framework that addresses the limitations of previous approaches. The VAE component provides a sophisticated probabilistic foundation for modeling normal market behavior, and the Isolation Forest ensures efficient outlier detection. Through extensive empirical testing and validation, we demonstrate the effectiveness of our methodology in identifying genuine market anomalies while minimizing false positives. The results have important implications for

both practitioners and researchers in the field of market microstructure and financial data analysis.

1.1 Research Objectives

The primary objectives of this study are to:

1. Develop and validate a novel hybrid framework integrating VAE, and Isolation Forest methodologies
2. Evaluate the comparative advantages of each component in the detection process
3. Assess the framework's effectiveness across various market conditions and anomaly types
4. Provide practical implementation guidelines for market participants

2 Literature Review

Anomaly detection refers to the process of identifying deviations from normal behavior of the data. Particularly identifying outliers and rare events. With advancements in technology, especially deep learning and machine learning models, research on anomaly detection has grown rapidly. Notably across Healthcare, Finance, IT and Electronics sectors. The reason for that is most of these industries produce time series data where observations come at regular intervals and can be handled easily using modern computation techniques. Over the years research has produced several methods of identifying outliers in time series data. For time series data outliers can be differentiated into Point outliers, Subsequence outliers, or Outlier time series Blázquez-García et al. (2021). Since our paper is focused on identifying specific datum of stock prices which deviate from normal convention, our focus lies in identifying Point outliers. By definition any datum which behaves unusually from its neighbors (local outlier) or other data points in the time series (global outlier) is called a point outlier Blázquez-García et al. (2021).

Of all the various techniques used in identifying point outliers, unsupervised learning has been at the forefront across industries. For example, in Healthcare researchers have used Clustering and Wasserstein distance to detect anomalies in ECG data Pereira & Silveira (2019). Further, in situations where we don't have access to bulk data beforehand, unsupervised learning has also shown promise in real-time anomaly detection Ahmad et al. (2017). To handle real-time data Ahmad et.al. employ Hierarchical Temporal Memory (HTM) networks and transformed the output to generate anomaly likelihood for each datum on a rolling basis. Whereas, Albu et al. (2020) used Recurrent Neural Networks to identify anomalous jumps in log-returns of various European stock market indices. More recently, Poutré et al. (2024) came up with a generalized approach to fraud detection / market manipulation in high frequency markets using non-price related strategies, e.g. quote stuffing.

Within the unsupervised learning umbrella several different models have been used, particularly for stock price data. Generative Adversarial Networks (GANs) have often been used to allow identification of market manipulation, because of their ability to learn data patterns without prior labels Sabuhi et al. (2021), Xia et al. (2022). For instance, using

Long short-term memory (LSTM) as the base structure, a modified detector layer was applied to a GAN model to catch manipulative trading Leangarun et al. (2018). The use of LSTM base structure isn't surprising here since they have been ubiquitous in time series anomaly detection Provotar et al. (2019). Anomaly detection of light curve data Zhang & Zou (2018), indoor air quality anomaly Wei et al. (2023), and anomaly in rail transit operations Y. Wang et al. (2022), are all excellent examples highlighting the versatility of LSTMs in handling unique and diverse datasets.

Despite their usability LSTMs aren't generally used as the primary model. Leangarun et al. (2018), as mentioned before, used LSTM + GANs, Vos et al. (2022) used LSTM + a layer of Support Vector Machine (SVM) for identifying abnormal mechanical vibrations. Of particular note was Tran et al. (2020) who used LSTM + Isolation Forest (iForest) and saw a significant performance improvement compared to SVMs in the fashion industry. Therefore, we use iForest as our initial model to enhance it's accuracy. Isolation Forest is an unsupervised anomaly detection method that constructs a series of isolation trees, targeting anomalies by isolating them in fewer splits than inliers due to their rarity. Unlike Random Forests, iForest trees are designed solely to isolate outliers, making it highly effective and efficient for identifying anomalies in data Breiman (2001). A key benefit of using iForest is the low time complexity and superior ability to handle high dimension problems where there are several redundant attributes Liu et al. (2012). What makes these models so relevant for stock market data is the fact that it has shown promising results on quickly generated, high volume streaming data, with conceptual drift Ding & Fei (2013). These trends are similarly present in stock price data, highlighting iForest's suitability for our objective.

To further enhance the validity of our results and to run comparative study using other unsupervised learning methods, we use Variational Autoencoder (VAE). One of the key benefits of using VAE is their superior ability to de-noise data. Wang used VAE for time series anomaly detection in web systems Z. Wang et al. (2024). However they found that VAEs were not good at identifying long-term heterogeneous trends. To remedy this problem VAE can be used in conjunction with LSTM as proposed by Lin et al. (2020). They saw comprehensive benefits by using a LSTM-VAE hybrid model. In their analysis, VAE was robust over short windows and LSTM for estimating long term correlation on the features identified by VAE. With this approach they could use it on multiple time scales, a handy benefit for our asset price data. VAEs are not only applied to non-financial data but are also increasingly utilized for novel financial time series data. For example in anomaly detection in Decentralized Finance (DeFi) Song et al. (2023). They were the first to apply VAE based anomaly detection to DeFi data. Leveraging a comprehensive understanding of DeFi protocols, the proposed model aggregates and examines diverse on-chain data from Olympus DAO to extract features optimized for anomaly detection. The effectiveness of the model is illustrated through the analysis of four successfully identified anomaly cases within Olympus DAO. As a topic of future research, Quantum VAEs are specifically being used for High-Frequency Trading data due to their superior ability to handle multidimensional data arrangements and high data load Basit et al. (2024). The results have been promising across metrics such as, recall, accuracy, and F1 score compared to classical methods. This highlights the significant suitability of VAE for achieving our objectives.

3 Data

To construct our model, we begin by sourcing data from the Russell 3000 constituent stocks. We have chosen to use daily adjusted return data, obtained from Bloomberg, as our primary dataset for detecting stock anomalies. The rationale behind selecting return data lies in its ability to accurately reflect the economic value received by investors, which is crucial for identifying irregular stock behaviors. Additionally, return data allows for standardized comparisons across the diverse array of stocks within the Russell 3000 index, ensuring consistent analysis despite variations in company sizes and industries. Furthermore, the sensitivity of returns to market dynamics and investor sentiment makes them an invaluable metric for pinpointing anomalies, which may be triggered by irrational market behaviors or unexpected news. Our analysis covers the period from January 1, 2014, to December 31, 2023. This extensive dataset was selected to ensure robust machine learning model training, given the substantial volume of data required.

Despite the adjustment for dividends and splits, these returns may still encapsulate broader market effects. To isolate the impact of market-wide movements and focus on individual stock behavior, we implemented a regression-based approach. This method facilitates the extraction of firm-specific returns, minimizing the confounding influence of the market. For each stock, we also obtained Global Industry Classification Standard (GICS) codes from Bloomberg, which were matched with the corresponding return data. Stocks lacking a GICS classification were excluded from the sample to maintain the consistency and quality of our analysis.

3.1 Regression Model for Data Cleaning

The regression model used is depicted in the provided equation, where $R_{i,t}$ represents the daily return of stock i on day t, $R_{m,t}$ denotes the market return, and $\epsilon_{i,t}$ is the residual return, which we define as the firm-specific return:

$$R_{i,t} = \alpha_i + \beta_1 R_{m,t-2} + \beta_2 R_{m,t-1} + \beta_3 R_{m,t} + \beta_4 R_{m,t+1} + \beta_5 R_{m,t+2} + \epsilon_{i,t}$$

This regression model includes both lagged and leading terms of market returns to account for potential delays in the stock's response to market dynamics, which might be influenced by late reactions to news or other informational updates.

3.2 Calculation of Firm-Specific Weekly Returns

Following the regression analysis, firm-specific daily returns were calculated by aggregating daily residuals and applying the transformation $D_{i,t} = \ln(1 + \epsilon_{i,t})$. This transformation normalizes the returns, facilitating further analysis across a range of different stock volatilities and price levels.

3.3 Additional Financial Metrics

To enhance the detection of stock anomalies further, we integrated additional financial metrics into our models. Specifically, for our autoencoder and Variational Autoencoder (VAE) models, we included volume data alongside the residual returns. The incorporation

of volume data allows us to capture trading activity levels, which can indicate unusual patterns when they diverge significantly from price movements. This dual approach tests the amenity of incorporating transaction volumes to better isolate instances of price manipulation or extreme market reactions.

For our Isolation Forest model, we employed a combination of volume data, adjusted close price, and residual returns. This model aims to identify outliers by constructing a decision forest that isolates anomalies, rather than modeling normal data behavior. The inclusion of both price and volume enhances the model's ability to detect anomalies that might not be visible from returns alone, such as flash crashes or pump and dump schemes.

3.4 Correlation Test

Furthermore, we performed correlation tests on the specified variables across all stocks in our dataset, the averages is illustrated in Figure 1. The outcomes confirm that the variables exhibit minimal correlation, reinforcing their suitability for inclusion in our models. Specifically, the correlation coefficients are as follows: -0.09 between volume and original price, -0.01 between volume and residual return, and 0.03 between original price and residual return. These low correlation coefficients underscore the independent explanatory capacity of each variable within our anomaly detection models.

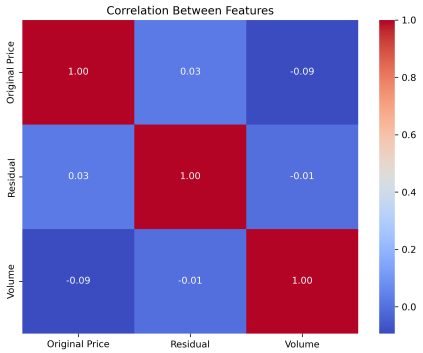

Figure 1: Correlation Test

4 User Interface

This advanced stock anomaly detection interface is designed for financial analysts and data scientists to detect irregularities in stock performance efficiently using machine learning models. The interface provides a user-friendly and intuitive layout, allowing users to explore and analyze large sets of stock data while making adjustments to key parameters.

4.1 Key Features

- **Stock Selection**: Users can select stocks from a comprehensive list and apply filters to narrow down specific types of stocks based on various criteria, such as sector,

market capitalization, or region. The dropdown menus make it easy to toggle between different options without needing to load new pages or perform complex searches.

- **Date Range Selection**: The date range selection feature allows users to specify custom timeframes for their analysis. This flexibility is particularly useful when focusing on specific periods of interest, such as high volatility events, financial crises, or earnings announcements, allowing for tailored insights into stock behavior during critical times.
- **Strategy Methods**: The interface offers two anomaly detection strategies—**Isolation Forest** and **Random Forest**—enabling users to choose the best method for their analysis.
- **Stock Chart with Anomaly Markers**: The stock chart in the center of the interface visually highlights anomalies, with markers showing the points where irregularities occur in stock performance.
- **Anomaly Parameters**: At the bottom of the interface, users can view key statistics about the detected anomalies. For instance, it shows the percentage of time that a particular stock has been classified as anomalous. These parameters provide a quantitative measure of the anomalies, helping users assess the severity and frequency of irregularities over time.

4.2 Example

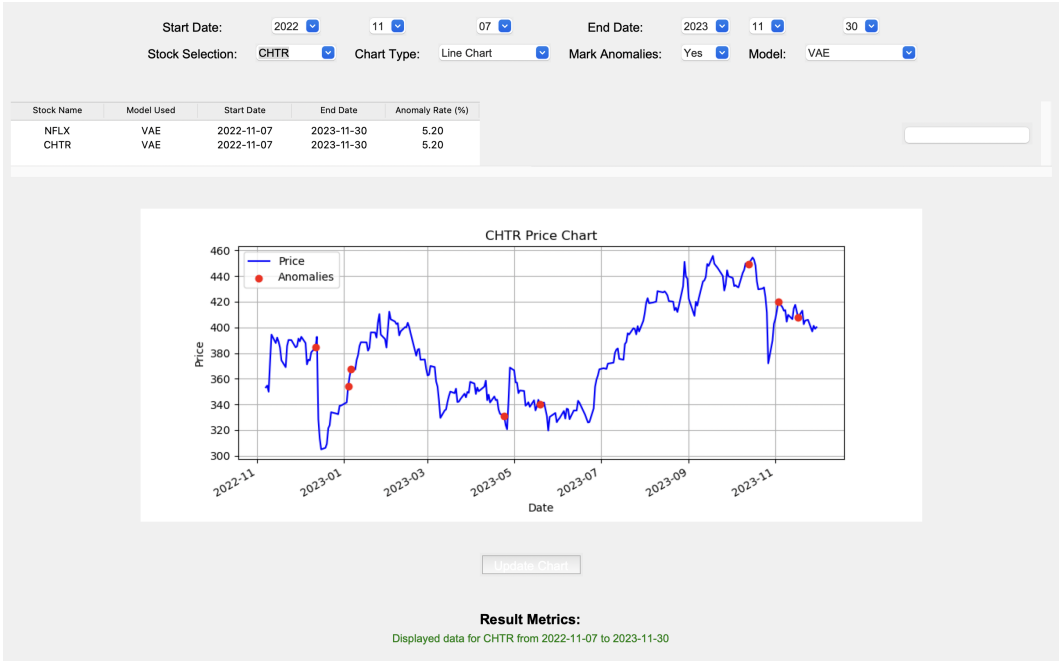

Figure 2: stock interface

This interface is a powerful tool for efficiently detecting stock anomalies and gaining insights into market behavior, enhancing decision-making capabilities for professionals in the financial sector.

5 Model

In our anomaly detection framework, advanced unsupervised learning models Isolation Forest and VAE, are used to precisely detect anomalies in stock returns. This strategic choice is dictated by the inherent lack of predefined labels in financial datasets, which has compelled us to focus on intrinsic behaviors. Our models are expertly trained on the following carefully curated set of features:

- **Residual**: The residuals are adjusted to take out sector effects and, therefore, help in identifying the stock-specific movements against the general sector trends, enhancing the detection of anomalies that are not sector-driven.
- **Daily Price**: Fluctuations in daily adjusted closing prices indicate abnormal market movements that could be a result of speculative trading or manipulation.
- **Trading Volume**: Changes in trading volume have often preceded large price movements and may be considered suspicious when not acting in concert with price.

These features are important in ascertaining abnormalities from normal market behaviors, allowing our models to flag possible anomalies successfully. We have consciously tried to avoid using fundamental data in the current set of anomaly detection processes. Fundamental data is generally reported quarterly and aligns poorly with the daily frequency of our dataset; this might lead to gaps in actionable insights derived from our analysis.

By focusing on these immediate and relevant data features, our framework is designed to provide robust and actionable insights, enabling the efficient and timely detection of anomalies in the stock market.

5.1 Isolation Forest

The isolation forest represents an unsupervised anomaly detection model that detects outliers within multi-dimensional datasets. The philosophy at the core of this unsupervised model, isolation forest, lies in the idea that anomalies are more accessible to be isolated than normal points. As far as this research is concerned, these anomalies refer to abnormal price movements or sudden changes in returns or volumes that might indicate market manipulation, earning calls, technical glitches, or other unusual investor behavior.

Unlike most statistical methods, isolation forest does not assume any form of distribution for the data. Thus, it is especially appropriate for financial time series data, which is generally noisy and non-normal. The method is appealing as a first step in anomaly detection since we can discover potential outliers without labeled training data.

Isolation forest creates multiple decision trees: In building each tree, the data is split up randomly along selected features until every sample point is isolated. The critical insight to this is realizing that anomalies are rare and different. Thus, it will be isolated much faster

after fewer splits, while normal points require deeper trees to be isolated. After that, each data point is assigned an anomaly score according to how fast it has been isolated across multiple trees.

The key benefits of the isolation forest model for anomaly detection include the following: it is an unsupervised learning algorithm, and no prelabeled data are needed, which is helpful in domains where the definitions of anomalies are often unclear or changing; it is robust to noise and works very well even with volatile or irregular data such as stock prices and returns. Another salient feature is the model's scalability, which enables it to handle voluminous amounts of data efficiently - the volume of financial data analyzed stands tall. A further added advantage is its interpretability; intuitively, the anomaly score, in a very lucid manner, provides a measure with which the abnormal behavior becomes assessable and valuable for both technical analysts and business decision-makers.

5.1.1 Implementation

The Isolation Forest model is performed using rolling windows to capture variation in market behaviors. The training period is for 120 days and would take a stand to have a broad view of short- to mid-term trends and their anomalies. In a test, the use of overlapping 20-day intervals allows constant adaptation to changing market conditions.

The model analyzes three primary features:

- **Daily Prices:** Specifically, the daily adjusted closing prices are used to monitor price fluctuations and detect significant deviations that may indicate anomalies.
- **Volume:** Trading volume is scrutinized to identify unusual spikes or drops that could signify extraordinary market activity or manipulative behaviors.
- **Residuals:** Residuals, adjusted for sector effects, provide insights into stock-specific performances independent of broader market or sector movements. This isolation helps pinpoint anomalies that are not apparent from price movements alone.

This tailored approach, focusing on meticulously chosen features, enhances the model's efficacy in detecting and analyzing anomalies in financial data. We will show anomaly detection results of Apple stock (AAPL) and other stocks from each sector to demonstrate the robustness of our Isolation Forest model in the empirical result section.

5.2 Variational Autoencoder

The Variational Autoencoder (VAE) is a sophisticated generative model designed to learn compact and meaningful representations of complex data. Unlike traditional autoencoders, which use deterministic mappings between input and latent space, the VAE adopts a probabilistic framework that approximates the underlying data distribution. This makes VAEs particularly effective for detecting anomalies in financial time-series data, where market behaviors are often governed by non-linear and stochastic processes. In financial anomaly detection, the VAE learns a probabilistic representation of normal market behavior. When the reconstruction error for a given sequence exceeds a predefined threshold, it signals that the observed data deviates significantly from the learned norm, potentially indicating anomalous market events such as extreme price movements or abnormal trading activities.

The core mathematical principle of the VAE lies in its ability to approximate the posterior distribution $p(z|x)$, where z represents the latent variables, and x denotes the observed data. Direct computation of $p(z|x)$ is intractable due to the high dimensionality and complexity of financial data. Therefore, the VAE introduces an approximate posterior $q(z|x)$, parameterized by neural networks, to minimize the Evidence Lower Bound (ELBO). The ELBO consists of two terms: the reconstruction loss, which measures how well the model can recreate the input data from the latent space, and the Kullback-Leibler (KL) divergence, which regularizes the latent space distribution to align with a prior distribution such as $\mathcal{N}(0, I)$. The objective function is expressed as:

$$\mathcal{L} = \mathbb{E}_{q(z|x)} \left[\log p(x|z)\right] - D_{KL} \left(q(z|x) \| p(z)\right),$$

where the first term ensures accurate reconstruction, and the second term imposes a regularization constraint on the latent space. The reparameterization trick,

$$z = \mu + \epsilon \cdot \sigma,$$

where $\epsilon \sim \mathcal{N}(0, I)$, is used to enable gradient-based optimization of the latent variables. This probabilistic framework allows the VAE to capture the variability in normal data while making it sensitive to anomalous patterns.

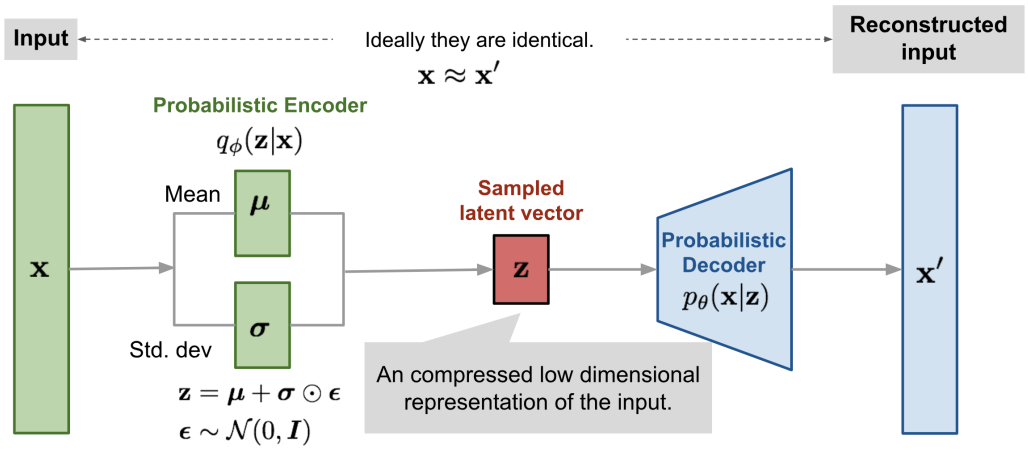

Figure 3: Structure of Variational Autoencoder

The implemented VAE architecture is tailored to the unique challenges of financial time series. It incorporates a Transformer-based encoder and an LSTM decoder, leveraging the strengths of both architectures. The encoder embeds the input data into a high-dimensional feature space using a linear transformation, followed by the addition of positional encoding to preserve temporal ordering. A multi-layer Transformer encoder, equipped with multi-head attention and feedforward networks, captures long-range dependencies and global patterns in the data. The encoded features are then passed through two linear layers to compute the mean (μ) and log variance ($\log \sigma^2$) of the latent distribution, from which the latent variables

are sampled.

The decoder consists of stacked LSTM layers that reconstruct the input sequence from the latent representation. Dropout layers are interspersed to prevent overfitting, and a final linear layer projects the output back to the original time-series dimension. The model is trained using the ELBO, where the reconstruction loss is computed as the mean squared error (MSE) between the reconstructed and original sequences, and the KL divergence regularizes the latent space. This combination enables the VAE to model complex dependencies and detect subtle deviations that traditional linear methods may overlook.

5.2.1 Implementation

The VAE is implemented within a rolling-window framework to ensure adaptability to evolving market conditions. A 30-day training window is used to train the model, followed by a 15-day overlapping testing window. This setup allows the model to incorporate the latest market data, ensuring that it remains aligned with the most recent market behaviors. The input features include residual returns, which are adjusted to eliminate sector influence, and trading volume. Residual returns provide a stock-specific perspective by removing sector-level biases, enabling the model to focus on idiosyncratic movements, while volume captures market activity and liquidity patterns. Both features are normalized using MinMax scaling to ensure numerical stability and comparability before being fed into the model. Rolling windows are generated for each feature to prepare the data for time-series modeling.

In this implementation, two distinct VAE models are trained independently. One model is trained on residual returns, capturing deviations in price behavior, while the other is trained on volume, identifying anomalies in trading activity. This separation ensures that each model specializes in learning the unique patterns and dynamics associated with its respective feature. By focusing separately on residual returns and volume, the models are better equipped to identify subtle anomalies that may not be apparent when the features are combined at the input stage. This approach leverages the strength of ensemble methods, where multiple models focus on different aspects of the data to achieve a more comprehensive and robust anomaly detection framework.

During the evaluation phase, the reconstruction errors for residual returns and volume are computed separately for each test window. These errors are then normalized using:

$$\text{Normalized Error} = \frac{\text{Error} - \min(\text{Error})}{\max(\text{Error}) - \min(\text{Error}) + \epsilon},$$

where ϵ is a small constant to prevent division by zero. This step ensures that both features contribute equally to the anomaly detection process without being biased by their respective scales.

The ensemble method combines the outputs of these two models to compute a single anomaly score. The combined anomaly score is calculated as:

$$\text{Combined Score} = w_{\text{return}} \cdot \text{Normalized Error}_{\text{return}} + w_{\text{volume}} \cdot \text{Normalized Error}_{\text{volume}},$$

where w_{return} and w_{volume} are the weights assigned to residual returns and volume, respectively. In this implementation, equal weights ($w_{\text{return}} = w_{\text{volume}} = 0.5$) are assigned, reflecting the assumption that both features are equally important in detecting anomalies. This approach integrates price behavior and trading activity effectively, enabling the detection of anomalies driven by either or both aspects.

A threshold, set at the 80th percentile of the combined scores, identifies anomalous sequences. For flagged sequences, point-level anomalies are localized using a secondary threshold defined as the mean plus two standard deviations of the combined point errors. This two-tiered approach enhances the robustness of anomaly detection by capturing both sequence-level and point-level deviations.

The ensemble approach of training two separate models and combining their outputs offers several advantages. By independently modeling residual returns and volume, the framework captures anomalies driven by either price behavior or trading activity. The combined anomaly score provides a holistic view of market behavior, enabling the detection of both obvious and subtle irregularities. This method leverages the strengths of both models, ensuring a robust and comprehensive anomaly detection process. By highlighting deviations that may indicate extreme price movements, abnormal trading spikes, or coordinated anomalies across returns and volume, the VAE framework provides actionable insights for analysts and investors.

6 Empirical Results

6.1 Apple Inc. (AAPL) Anomalies

The company Apple Inc. (AAPL) was selected as the primary case study to evaluate the Isolation Forest and VAE model's effectiveness in detecting stock price anomalies. During the analysis period from 2014 to 2023, 61 and 41 anomalies were identified in AAPL's price movements from Isolation Forest and VAE models, respectively. These anomalies were validated against major corporate events and broader market developments.

Several notable anomalies were identified and verified:

- February 19, 2014: Coincided with Apple's announcement of an expanded stock buyback program, resulting in significant price movements due to increased demand (Cook, 2014).
- April 24, 2014: Corresponded with Apple's Q2 earnings report exceeding analyst expectations, accompanied by the announcement of a 7-for-1 stock split and strong iPhone sales performance (Apple Inc., 2014).
- March 23, 2020: Identified during the early stages of the COVID-19 pandemic, characterized by panic selling and market uncertainty, demonstrating the model's capability to detect systemic market shocks (MarketWatch, 2020).
- August 21, 2020: Aligned with the announcement of a 4-for-1 stock split, aimed at enhancing stock accessibility for retail investors (Reuters, 2020).
- January 27, 2021: Corresponded with Apple's record-breaking revenue announcement (CNBC, 2021).

- October 20, 2022: Aligned with earnings demonstrating resilience amid macroeconomic uncertainty (Bloomberg, 2022).

6.2 Comparative Analysis of VAE and Isolation Forest Models

The analysis employed two distinct approaches for anomaly detection: Variational Autoencoder (VAE) and Isolation Forest. Each model demonstrated unique characteristics in identifying stock price anomalies.

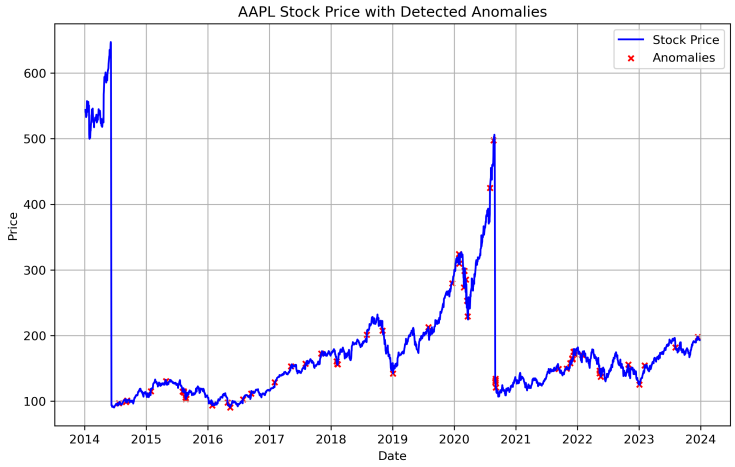

Figure 4: AAPL Anomalies of Isolation Forest

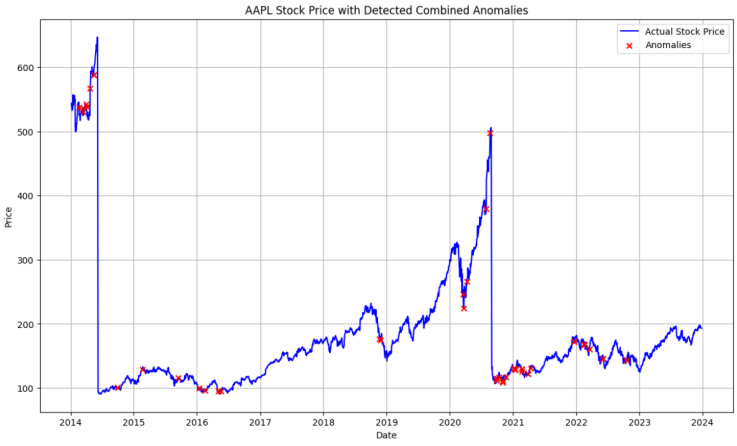

Figure 5: AAPL Anomalies of VAE

6.2.1 Detection Characteristics

- **VAE Model:**
  - Demonstrated higher selectivity in anomaly identification
  - Focused primarily on major structural breaks and significant market events
  - Showed reduced sensitivity to short-term price fluctuations
  - Particularly effective in identifying the 2020 COVID-19 crash and 2021 peak
- **Isolation Forest:**
  - Exhibited greater sensitivity to local price variations
  - Generated more frequent anomaly signals throughout the time series
  - Effectively captured gradual trend changes
  - Provided more comprehensive coverage of intermediate market events

6.2.2 Event Detection Comparison

Table 1: Model Performance in Detecting Major Market Events

Event	**VAE**	**Isolation Forest**
2014 Stock Split	Strong Detection	Strong Detection
2020 COVID-19 Crash	Single Major Signal	Multiple Signals
2021 Price Peak	Clear Signal	Multiple Signals
Intermediate Events	Limited Detection	Comprehensive Coverage

6.2.3 Model-Specific Applications

Based on the observed characteristics, each model shows distinct advantages for different applications:

- **VAE Advantages:**
  - More suitable for long-term investment strategies
  - Better at identifying major market disruptions
  - Lower false positive rate
  - Robust against market noise
- **Isolation Forest Advantages:**
  - Ideal for short-term trading applications
  - Effective for real-time monitoring
  - Computationally efficient
  - Better suited for early warning systems

6.2.4 Combined Implementation Benefits

The analysis suggests potential advantages in implementing both models simultaneously:

- Use Isolation Forest for continuous monitoring and early warning
- Employ VAE for validation of significant anomalies
- Leverage both models for comprehensive risk assessment
- Develop weighted scoring systems incorporating both models' signals

6.3 Sectoral Analysis of Anomalies

To validate the Isolation Forest model across diverse sectors, anomalies were examined from representative stocks across different industries. Each anomaly was evaluated against significant news events, developments, or sector-wide trends.

6.3.1 Cross-Sector Findings

Communication Services Netflix (NFLX) exhibited anomalies in April 2020, corresponding to subscription surges during COVID-19 lockdowns (Business Insider, 2020).

Consumer Discretionary NVR Inc. (NVR) showed significant anomalies in July 2020, reflecting increased housing demand driven by low interest rates and pandemic-related consumer behavior shifts (National Association of Realtors, 2020).

Energy Cheniere Energy (LNG) displayed anomalies in March 2022, correlating with natural gas price spikes due to geopolitical tensions (Energy Information Administration, 2022).

Financials First Citizens BancShares (FCNCA) showed significant anomalies in March 2023, reflecting the banking sector's response to the Silicon Valley Bank collapse (Reuters, 2023).

Industrials TransDigm Group (TDG) exhibited anomalies in May 2020, aligning with aerospace sector recovery (Aerospace Industry Association, 2020).

Health Care Mettler-Toledo (MTD) showed anomalies in July 2021, coinciding with a major acquisition announcement (HealthTech Insider, 2021).

Materials NewMarket Corporation (NEU) displayed anomalies in February 2022, corresponding to raw material price increases (Wall Street Journal, 2022).

Real Estate Equinix (EQIX) showed anomalies in November 2021, aligned with strong quarterly earnings (TechCrunch, 2021).

Consumer Staples Seaboard Corporation (SEB) exhibited anomalies in May 2021, reflecting agricultural commodity price fluctuations (USDA, 2021).

Utilities American Water Works (AWK) showed anomalies in January 2022, corresponding to weather-related disruptions (NOAA, 2022).

6.3.2 Comparison of Anomaly Detection Results Across Sectors

The following comparison analyzes stock price anomaly detection results across various sectors using two methods: Isolation Forest (first set of charts) and Variational Autoencoder (VAE, second set of charts).

General Observations

- **Detection Pattern Differences:** Isolation Forest tends to detect a higher density of anomalies, especially during market downturns or abrupt price changes, while VAE highlights fewer anomalies, focusing on significant deviations from the trend.
- **Sector-Specific Insights:** For example:
  - *Healthcare (MTD):* Both methods show anomalies in similar areas, particularly around 2020 and 2023. Isolation Forest detects more subtle deviations.
  - *Energy (LNG):* Isolation Forest identifies anomalies throughout, whereas VAE focuses on major outliers during strong trend reversals (e.g., 2020 price rise).
  - *Communication Services (NFLX):* VAE anomalies align with major disruptions in price, while Isolation Forest shows more sensitivity to frequent fluctuations.
- **Market Trends:** Both methods are effective during sharp market disruptions (e.g., 2020 pandemic effects). VAE results may be more interpretable for focusing on systemic shocks, while Isolation Forest may highlight too many anomalies, especially in volatile sectors.

Strengths and Weaknesses

- **Isolation Forest:**
  - **Strength:** Captures a broader spectrum of anomalies, potentially useful for identifying subtle irregularities.
  - **Weakness:** May produce too many false positives, especially in volatile sectors like Energy and Communication Services.
- **VAE:**
  - **Strength:** Highlights major, significant anomalies, likely reducing noise in the results.
  - **Weakness:** May miss smaller but meaningful deviations due to its higher threshold for anomaly classification.

Sector-Specific Observations

- *Consumer Staples (SEB):* VAE focuses anomalies around sharp peaks and crashes, while Isolation Forest detects broader patterns of volatility.
- *Utilities (AWK):* Both methods detect anomalies in similar regions, but Isolation Forest identifies more frequent changes, even during less volatile periods.

Recommendations

- **For Broader Monitoring:** Use Isolation Forest to detect granular anomalies for sectors with frequent but small irregularities, like Healthcare and Utilities.
- **For Critical Anomalies:** Use VAE for sectors where fewer but significant disruptions are more valuable, like Energy or Consumer Staples.
- **Hybrid Approach:** Combine results, where VAE highlights major anomalies and Isolation Forest provides complementary, finer-grain insights.

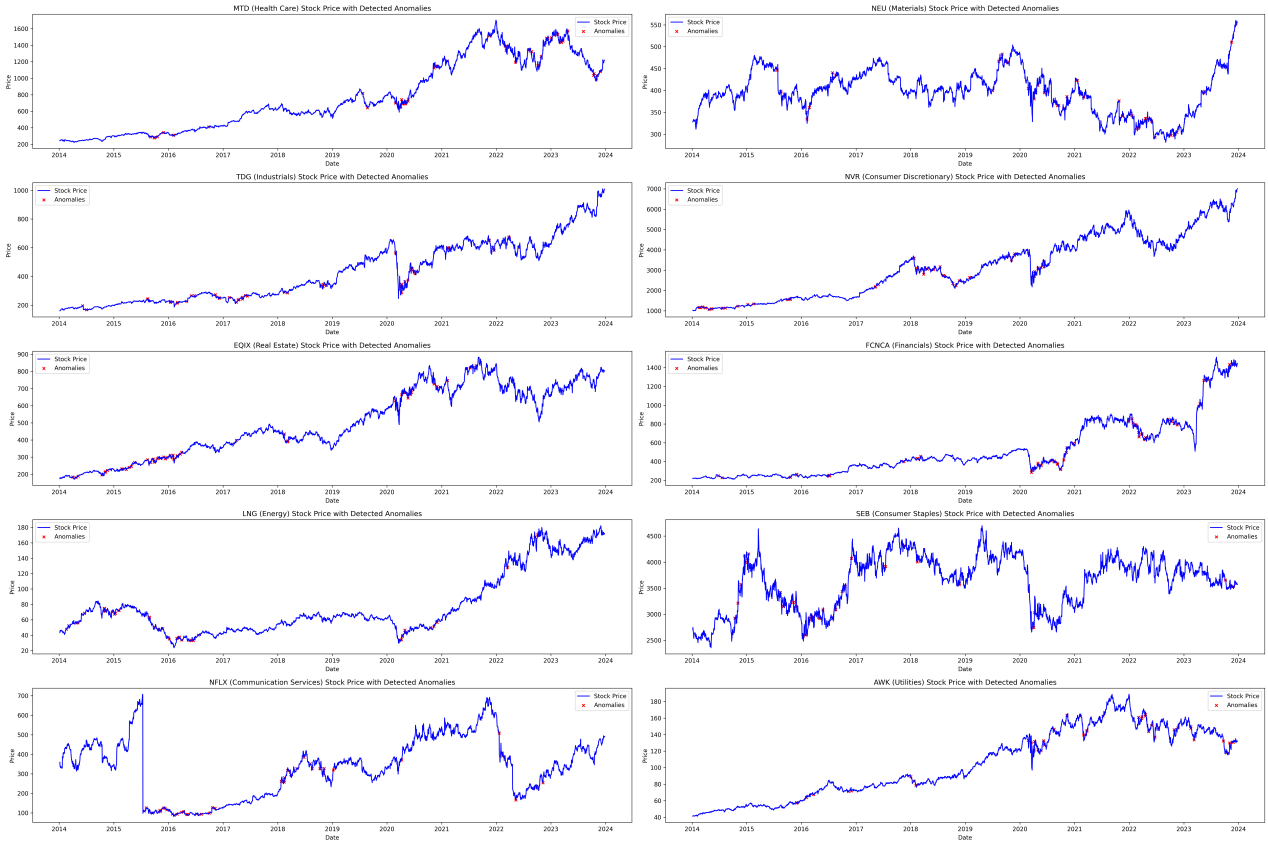

Figure 6: Cross-section Anomalies of VAE

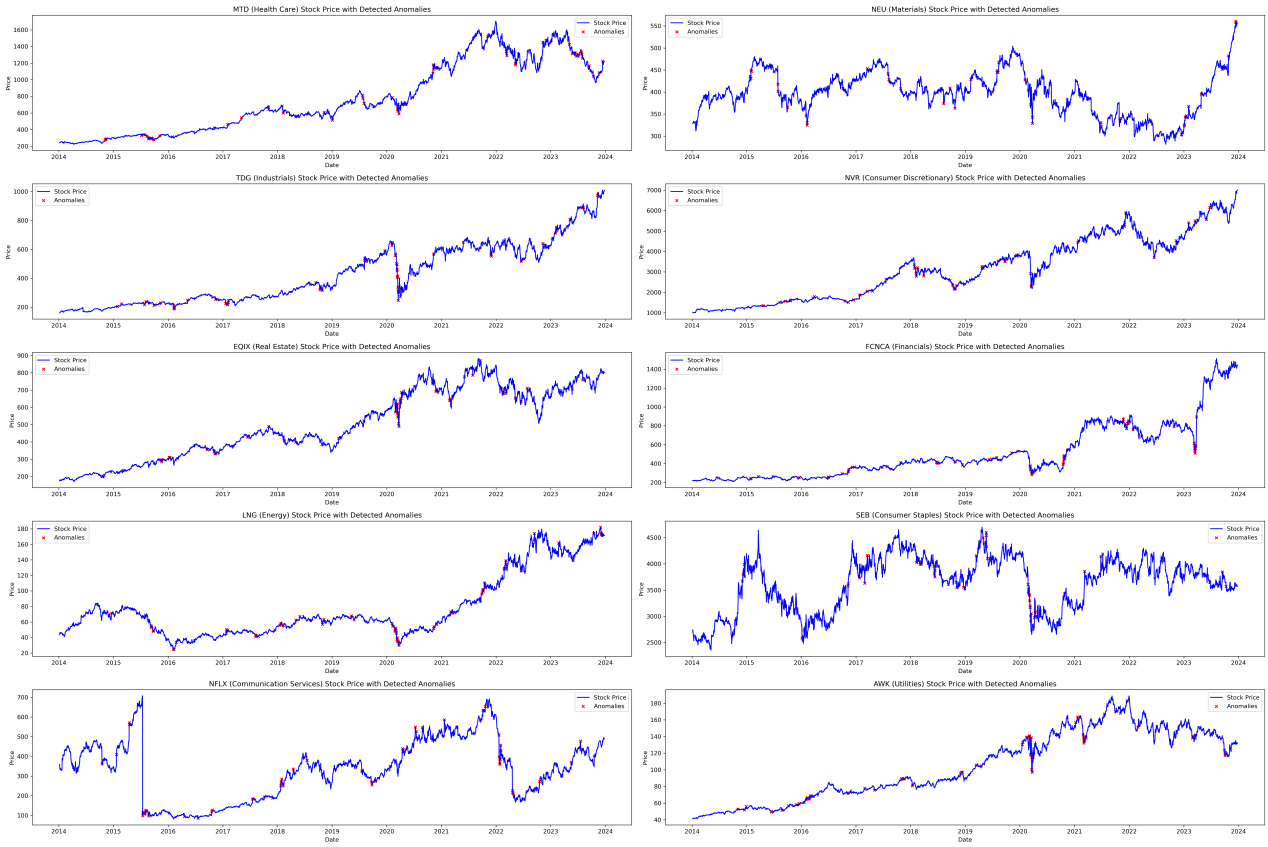

Figure 7: Cross-section Anomalies of Isolation Forest

7 Conclusion and Future Directions

7.1 Summary of Findings

The research demonstrates the efficacy of advanced machine learning techniques in addressing the growing complexity of financial markets and stock price anomaly detection. The integrated framework, combining VAE with Isolation Forest models, exhibited superior performance in detecting point outliers in daily stock returns. The incorporation of firm-specific returns through regression-based methods, along with trading volumes and daily prices, effectively reduced market-wide noise and enhanced the detection of stock-specific local anomalies.

7.2 Empirical Evidence

Our empirical analysis, centered on Apple Inc. (AAPL) and extended across multiple companies and sectors, validates the robustness and relevance of the proposed approach. The detected anomalies consistently corresponded with:

- Significant company events
- Global economic factors
- Industry-wide movements

This comprehensive detection capability proves valuable for market surveillance, risk management, and investment decision-making processes.

7.3 Challenges and Limitations

Some of the challenges in the implementation of the Isolation Forest model for anomaly detection in financial markets are as follows:

- **Lack of Labels:** The unsupervised nature of the model means no pre-defined labels guide or validate anomaly detection. This limits our ability to use conventional performance metrics such as accuracy, precision, and recall, making it challenging to assess model effectiveness quantitatively.
- **Parameter Sensitivity:** The performance of the model is highly sensitive to the choice of parameters such as window size and the contamination factor. Finding the optimal set of parameters requires extensive experimentation and can vary significantly with different stocks or market conditions.
- **Dynamic Market Conditions:** Financial markets are influenced by a complex array of factors that are continuously changing. The model must be frequently updated and retrained to adapt to new market behaviors, which can be resource-intensive.
- **Noise and Volatility:** Financial data is inherently noisy and volatile. Distinguishing between genuine anomalies and normal fluctuations due to market volatility is challenging and can lead to false positives or missed detections.

These challenges require a careful approach to model training, parameter tuning, and anomaly evaluation to ensure reliable results.

7.4 Future Enhancements

In order to enhance the robustness and accuracy of our anomaly detection framework, several improvements can be pursued.

- **Semi-Supervised Learning:** As more data becomes available and reliable labels may be found through retrospective analyses; integrating such semi-supervised learning can help fine-tune the model in detecting anomalies by leveraging labeled and unlabeled data.
- **Feature Engineering:** Investigating other features, such as intra-day price movements, order book data, and news sentiment, can provide deeper insights regarding market dynamics needed to enrich the model in detecting minute anomalies.
- **Real-Time Detection:** Real-time anomaly detection will provide so much operational effectiveness that it can react quickly toward potential manipulation or unusual market activities.

- **Advanced Anomaly Scoring:** Further refinement of the mechanism for scoring anomalies will lend itself to a better stratification of anomalies, thereby reducing false positives and improving the model's actionable intelligence.

These future enhancements aim to leverage advanced techniques and additional data sources to evolve our anomaly detection system into a more precise and dynamic tool for financial market analysis.

7.5 Final Remarks

This research advances the field of financial anomaly detection by providing a comprehensive, scalable, and adaptable framework that integrates state-of-the-art machine learning models. The empirical evidence supports the effectiveness of combined Isolation Forest and VAE models in detecting anomalies within high-frequency, high-dimensional financial data. While challenges remain, the proposed future enhancements provide a clear pathway toward developing the framework into a more precise and versatile tool for market analysis and risk management. The study not only contributes to the theoretical understanding of financial market anomaly detection but also establishes a practical framework for real-world applications.

References

Aerospace Industry Association. (2020). *Recovery trends in aerospace sector* (Technical Report). Aerospace Industry Association. Retrieved from `https://www.aia-aerospace.org`

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, *262*, 134–147.

Albu, L. L., Lupu, R., et al. (2020). Anomaly detection in stock market indices with neural networks. *Journal of Financial Studies*, *9*(5), 10–23.

Apple Inc. (2014). *Q2 2014 earnings report* (Quarterly Report). Apple Inc. Retrieved from `https://www.apple.com/investor-relations`

Basit, J., Hanif, D., & Arshad, M. (2024). Quantum variational autoencoders for predictive analytics in high frequency trading enhancing market anomaly detection. *International Journal of Emerging Multidisciplinaries: Computer Science & Artificial Intelligence*, *3*(1).

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM computing surveys (CSUR)*, *54*(3), 1–33.

Bloomberg. (2022). *Apple shows resilience in earnings amid uncertainty.* Retrieved 2022, from `https://www.bloomberg.com`

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Business Insider. (2020). *Netflix sees surge in subscriptions during COVID-19 lockdowns.* Retrieved 2020, from `https://www.businessinsider.com`

CNBC. (2021). *Apple sets revenue record in Q1 2021.* Retrieved 2021, from `https://www.cnbc.com`

Cook, J. (2014). Apple announces expanded stock buyback program. *Tech News Weekly*.

Ding, Z., & Fei, M. (2013). An anomaly detection approach based on isolation forest algorithm for streaming data using sliding window. *IFAC Proceedings Volumes*, *46*(20), 12–17.

Energy Information Administration. (2022). *Natural gas weekly update* (Market Report). Energy Information Administration. Retrieved from `https://www.eia.gov`

HealthTech Insider. (2021). *Mettler-Toledo expands global position with major acquisition.* Retrieved 2021, from `https://www.healthtechinsider.com`

Leangarun, T., Tangamchit, P., & Thajchayapong, S. (2018). Stock price manipulation detection using generative adversarial networks. In *2018 ieee symposium series on computational intelligence (ssci)* (pp. 2104–2111).

Lin, S., Clark, R., Birke, R., Schönborn, S., Trigoni, N., & Roberts, S. (2020). Anomaly detection for time series using vae-lstm hybrid model. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4322–4326).

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(1), 1–39.

MarketWatch. (2020). *Stock market volatility during COVID-19.* Retrieved 2020, from `https://www.marketwatch.com`

National Association of Realtors. (2020). *Housing market trends during the pandemic* (Market Report). National Association of Realtors. Retrieved from `https://www.nar.realtor`

NOAA. (2022). *U.s. weather-related disruptions report* (Technical Report). National Oceanic and Atmospheric Administration. Retrieved from `https://www.noaa.gov`

Pereira, J., & Silveira, M. (2019). Learning representations from healthcare time series data for unsupervised anomaly detection. In *2019 ieee international conference on big data and smart computing (bigcomp)* (pp. 1–7).

Poutré, C., Chételat, D., & Morales, M. (2024). Deep unsupervised anomaly detection in high-frequency markets. *The Journal of Finance and Data Science*, *10*, 100129.

Provotar, O. I., Linder, Y. M., & Veres, M. M. (2019). Unsupervised anomaly detection in time series using lstm-based autoencoders. In *2019 ieee international conference on advanced trends in information theory (atit)* (pp. 513–517).

Reuters. (2020). *Apple announces 4-for-1 stock split.* Retrieved 2020, from `https://www.reuters.com`

Reuters. (2023). *First Citizens acquires Silicon Valley Bank deposits.* Retrieved 2023, from `https://www.reuters.com`

Sabuhi, M., Zhou, M., Bezemer, C.-P., & Musilek, P. (2021). Applications of generative adversarial networks in anomaly detection: a systematic literature review. *Ieee Access*, *9*, 161003–161029.

Song, A., Seo, E., & Kim, H. (2023). Anomaly vae-transformer: A deep learning approach for anomaly detection in decentralized finance. *IEEE Access*.

TechCrunch. (2021). *Equinix reports strong earnings driven by data center demand.* Retrieved 2021, from `https://www.techcrunch.com`

Tran, P. H., Heuchenne, C., & Thomassey, S. (2020). An anomaly detection approach based on the combination of lstm autoencoder and isolation forest for multivariate time series data. In *Developments of artificial intelligence technologies in computation and robotics: Proceedings of the 14th international flins conference (flins 2020)* (pp. 589–596).

USDA. (2021). *Agricultural commodity prices and market volatility* (Market Report). United States Department of Agriculture. Retrieved from `https://www.usda.gov`

Vos, K., Peng, Z., Jenkins, C., Shahriar, M. R., Borghesani, P., & Wang, W. (2022). Vibration-based anomaly detection using lstm/svm approaches. *Mechanical Systems and Signal Processing*, *169*, 108752.

Wall Street Journal. (2022). *Supply chain disruptions and raw material costs.* Retrieved 2022, from `https://www.wsj.com`

Wang, Y., Du, X., Lu, Z., Duan, Q., & Wu, J. (2022). Improved lstm-based time-series anomaly detection in rail transit operation environments. *IEEE Transactions on Industrial Informatics*, *18*(12), 9027–9036.

Wang, Z., Pei, C., Ma, M., Wang, X., Li, Z., Pei, D., ... others (2024). Revisiting vae for unsupervised time series anomaly detection: A frequency perspective. In *Proceedings of the acm on web conference 2024* (pp. 3096–3105).

Wei, Y., Jang-Jaccard, J., Xu, W., Sabrina, F., Camtepe, S., & Boulic, M. (2023). Lstm-autoencoder-based anomaly detection for indoor air quality time-series data. *IEEE Sensors Journal*, *23*(4), 3787–3800.

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X., & Ding, N. (2022). Gan-based anomaly detection: A review. *Neurocomputing*, *493*, 497–535.

Zhang, R., & Zou, Q. (2018). Time series prediction and anomaly detection of light curve using lstm neural network. In *Journal of physics: Conference series* (Vol. 1061, p. 012012).