Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 24, No. 3, July 2025, pp. 423∼438
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v24i3.4514

Leveraging Vector Quantized Variational Autoencoder for Accurate Synthetic Data Generation in Multivariate Time Series

Mohammad Diqi , Ema Utami , Kusrini , Ferry Wahyu Wibowo
Universitas Amikom Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

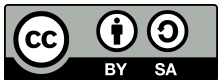
Received October 05, 2024
Revised January 24, 2025
Accepted May 10, 2025

Keywords:

Financial Market;
Multivariate Time Series;
Synthetic Data Generation;
Variational Autoencoder;
Vector Quantized.

ABSTRACT

This study addresses the challenge of generating high-quality synthetic financial time series data, a critical issue in financial forecasting due to limited access to complete and reliable historical datasets. The aim of this research was to compare the performance of the standard Variational Autoencoder and the Vector Quantized Variational Autoencoder (VQ-VAE) in generating synthetic multivariate time series data using the Adaro Energy Indonesia stock dataset. The VQ-VAE incorporates a discrete latent space to improve the structure and control of the data generation process, whereas the standard VAE utilizes a continuous latent space. **This research method** was based on the implementation of both models, followed by a quantitative evaluation using statistical metrics, including mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R^2 score. **This research showed** that the VQ-VAE outperformed the standard VAE in replicating the statistical characteristics of stock prices, as shown by lower error values and higher R^2 scores across all tested features. The discrete latent space of the VQ-VAE led to the generation of more structured and statistically consistent synthetic data. **The implications of these findings** suggest that the VQ-VAE model is highly suitable for financial forecasting applications and indicate the potential for future enhancements through integration with hybrid models, such as attention mechanisms or generative adversarial networks.



Corresponding Author:

Mohammad Diqi, 082264503889,
Department of Informatics Doctorate,
Universitas Amikom Yogyakarta, Yogyakarta, Indonesia,
Email: diqi@students.amikom.ac.id.

How to Cite:
M. Diqi, E. Utami, K. Kusrini, and F. W. Wibowo, ”Leveraging Vector Quantized Variational Autoencoder for Accurate Synthetic Data Generation in Multivariate Time Series”, *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 3, pp. 423–438, doi: 10.30812/matrik.v24i3.4514.

1. INTRODUCTION

Time series data plays a pivotal role in stock market analysis, providing the foundation for predicting future price movements [1]. By examining historical data points such as open, high, low, close, and volume, investors and analysts can identify patterns and trends that inform decision-making [2]. These predictions help market participants anticipate changes in stock prices, allowing for better timing of investments and risk management [3]. Companies also rely on time series data to forecast stock performance and adjust their strategies accordingly [4]. In an ever-evolving market, accurate predictions based on historical data are essential for staying competitive and informed [5]. However, access to high-quality historical data can be limited due to several factors [6]. Data may be incomplete, proprietary, or inaccessible for specific periods, leading to gaps in the dataset. In addition, companies may restrict access to their data, hindering broader research efforts. These limitations can pose significant challenges for those who need comprehensive data for accurate stock analysis [7]. One potential solution to this problem is synthetic data, which aims to duplicate the statistical properties of real data. It allows for the generation of datasets that mirror real market conditions without requiring access to restricted or incomplete data sources [8].

Generative models, such as Variational Autoencoders (VAEs), have become a powerful tool for creating synthetic data [9]. VAEs can generate new data from latent distributions that capture the underlying patterns of the original dataset [10]. By learning from the existing data, VAEs produce synthetic data that closely resembles real-world stock behavior, making them valuable in financial analysis and machine-learning applications [11]. This approach not only fills gaps in the dataset but also enhances the ability to train predictive models more effectively [12]. In the financial sector, the use of VAEs for generating realistic synthetic data has opened new possibilities for improving stock market forecasts and understanding market dynamics [13]. While standard VAEs are effective at generating synthetic data, they have limitations related to their continuous latent space representation [14]. The flexibility of continuous latent spaces can result in a lack of structure, making it challenging to control or organize the generated data effectively [15]. With clear boundaries or clusters, the latent space can produce data that is more interpretable and better aligned with the desired characteristics [16]. This often hinders the generation of well-organized synthetic datasets, especially in complex domains like stock markets [10]. For financial applications, this lack of control can diminish the precision and reliability of the data generated by standard VAEs [17].

The VAE has been extensively explored in various domains for synthetic data generation, with notable applications ranging from finance to bioinformatics. Sensitivity-based methods make VAE models more accessible to understand, especially when it comes to financial data, by showing how input variables affect the created synthetic tabular data [18]. This approach allows for global and local interpretations, addressing one of the major criticisms of the black-box nature of VAE models. In bioinformatics, VAEs have proven valuable in tasks like molecular design, multi-omics analysis, and biological sequence analysis, as they facilitate representational learning [19]. The strength of VAE in this field lies in its ability to generate synthetic data with high intra-class variation, which can be advantageous for understanding complex biological systems. However, the challenge of handling unlabeled data in bioinformatics remains a limitation, as this can impact the quality of the generated data. In more specialized applications, the combination of VAE with InfoGAN has shown promising results in generating semantically rich synthetic images for geospatial analysis. This hybrid model allows for better control over image features, although the computational demands of processing both pixel and feature conditions simultaneously can be a significant drawback [20]. Similarly, in financial applications, VAEs have been compared to CTGAN for generating credit data, where both models present unique advantages [21]. Despite these advances, the need for a gold standard for validating synthetic data in finance raises concerns about the reliability of these models. Lastly, in health-related fields, VAEs have been applied to augment eye-tracking data to support classification tasks with limited datasets [22]. While this technique enhances the available data, the synthetic outputs may only sometimes faithfully represent the real-world data, posing a challenge for model accuracy in medical applications.

Recent improvements in VAEs have had a significant impact on the fields of text-to-speech and anomaly detection [23]. This shows how flexible VAE techniques are when dealing with large datasets. A VAE model that uses vector quantization and an autoregressive prosody shows how important quantized latent spaces have been in making synthesized voices sound more natural. Although this model improves vocal naturalness, it struggles with unstable prosody variations across tokens [24]. In the same way, the Split Vector Quantized VAE (SVQ-VAE) improves the sampling efficiency in the latent space, which makes speech synthesis sound more natural [25]. However, this method requires intricate tuning of the latent space, complicating its practical deployment. When finding anomalies, VAEs with auto-regressive models in discretized latent spaces are better at finding and fixing image-based anomalies. This setup necessitates high computational power for image reencoding, which can be a limiting factor in resource-constrained environments [26]. Another innovative approach involves replacing vector quantization with finite scalar quantization (FSQ), which addresses the frequent issue of codebook collapse in traditional VQ-VAEs. Although FSQ shows promise in refining latent representations, it still requires further experimentation across broader domains to establish its efficacy and scalability [27]. These studies collectively illustrate the ongoing evolution of VAE technologies and growing applicability across diverse digital and

visual media.

Even though VAEs and VQ-VAEs have been shown to work well in areas like bioinformatics, geospatial imaging, and voice synthesis, significant gaps remain in understanding their effectiveness for financial time series data. Existing studies predominantly focus on isolated applications within specific domains [28], with limited exploration of the comparative performance of these models in synthesizing stock market data [9]. **This gap** is critical given financial datasets' unique challenges, such as high volatility, nonlinear market behavior, and the need for precise data representation. Furthermore, standard VAEs often struggle with generating well-structured synthetic data due to the inherent flexibility of their continuous latent space, which may lead to less control over data generation and reduced reliability in complex domains like finance. To address these **limitations**, this study systematically compares VAE and VQ-VAE in generating synthetic stock data, emphasizing the benefits of discrete latent spaces introduced by VQ-VAE. By focusing on essential metrics such as the R^2 score, mean absolute error (MAE), and mean squared error (MSE), the research highlights the strengths and weaknesses of each model in replicating the statistical properties of real stock data. This analysis underscores the ability of VQ-VAE to produce more structured and accurate synthetic data, addressing the weaknesses observed in standard VAEs when applied to financial time series data.

The **novelty** of this study lies in its application of VQ-VAE to the Indonesian stock market, specifically PT Adaro Energy Indonesia TBK (ADRO) data, a context that has received little attention in prior research. While previous works have explored VAEs in various fields, they often lack a focus on controlling and enhancing latent space representation for financial datasets. This research fills this gap by demonstrating how the discrete latent spaces of VQ-VAE can overcome limitations in traditional approaches, leading to improved data generation quality. It also contributes to the literature by providing a real-world case study, offering practical financial forecasting and modeling insights. **This research aims** to evaluate and compare the effectiveness of VAE and VQ-VAE in generating high-fidelity synthetic multivariate time series data for financial forecasting, specifically focusing on replicating the statistical characteristics of stock market data. In summary, this study addresses the limitations of prior works by proposing a robust framework for evaluating generative models in financial time series data. It **contributes** theoretically and practically by showing how VQ-VAE's structured latent space enhances data generation, ultimately paving the way for more reliable and precise synthetic data applications in finance.

2. RESEARCH METHOD

This section outlines the research methodology adopted in evaluating the performance of the VAE and VQ-VAE for generating synthetic multivariate financial time series data. This research method is structured into a clear flow comprising six key stages, as shown in Figure 1.

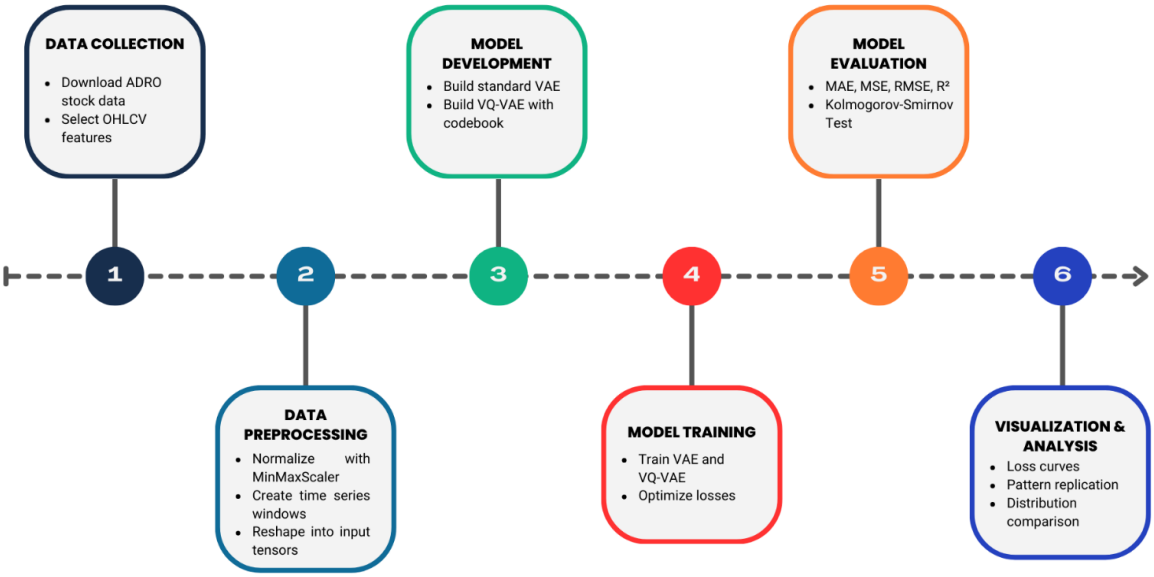


Figure 1. Research flow diagram for synthetic financial time series data generation using VAE and VQ-VAE

2.1. Dataset and preprocessing

The stock data for Adaro Energy Indonesia (ADRO) was obtained from Kaggle [29], spanning the period from July 16, 2008, to April 29, 2024. This dataset covers 3.803 trading day records and includes key financial features such as open, high, low, close, and volume, essential for analyzing stock market trends, as shown in Figure 2. To prepare the data for model training, we applied MinMaxScaler to normalize the values, ensuring they fall from 0 to 1. This normalization process is critical for improving the stability and performance of generative models like VAE and VQ-VAE. Scaling the data allows the models to process and learn from the dataset more effectively, leading to better synthetic data generation.

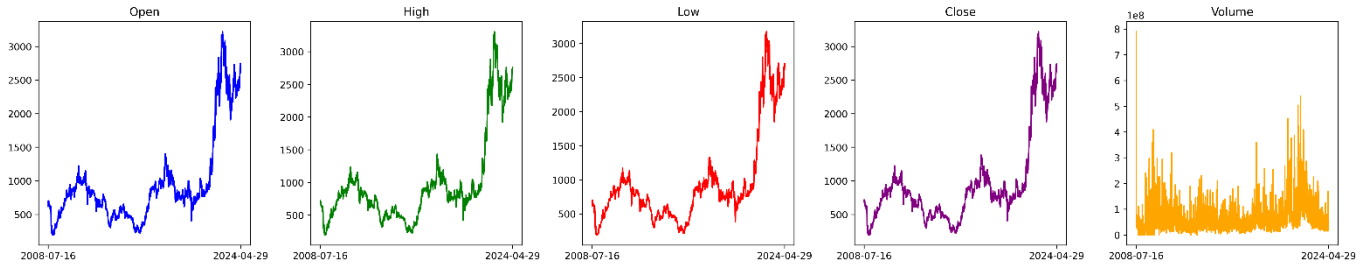


Figure 2. Dataset of PT adaro energy Indonesia TBK (ADRO)

The time series data was structured using a specified sequence length to prepare the data for neural network input, allowing the model to capture temporal dependencies between data points. Each sequence includes a fixed number of consecutive stock price observations, crucial for understanding trends over time. Once structured, the data was reshaped into a tensor format compatible with the neural network's architecture. This reshaping process organizes the data into a three-dimensional format, where each tensor represents a sequence of time steps, features, and batches. By transforming the data this way, we ensure the model can efficiently process and learn from the time series patterns.

2.2. Model Architecture

2.2.1. Standard VAE

The standard VAE is a generative model designed to learn a latent representation of input data to generate new, similar data. The architecture consists of two main components: the encoder and the decoder [19]. The encoder compresses the input data into a lower-dimensional latent space by producing two outputs: the mean μ and the log-variance $\log \sigma^2$ which together define a Gaussian distribution in the latent space. The reparameterization trick is applied to samples from this distribution, allowing for backpropagation during training [30]. Equation 1 obtains the latent variable z instead of directly sampling from the distribution. Where $\epsilon \sim N\ (0.1)$ is a random variable drawn from a standard normal distribution, and $\sigma = exp\left(\frac{\log \sigma^2}{2}\right)$

$$z = \mu + \sigma.\epsilon \tag{1}$$

The decoder utilizes the latent variable z to rebuild the original input data, striving to minimize the discrepancy between the original and recreated data. The loss function of a VAE integrates two components: reconstruction loss and Kullback-Leibler (KL) divergence. The reconstruction loss quantifies the decoder's efficacy in reconstructing the input, employing mean squared error (MSE) as calculated in Equation 2. where x is the original input and x is the reconstruction.

$$\text{Reconstruction Loss} = |x - \hat{x}|^2 \tag{2}$$

The second component, KL Divergence, encourages the latent distribution to remain close to a standard Gaussian distribution $N(0.1)$. It measures the divergence between the learned posterior distribution $q(z|x)$ and the prior $p(z)$, which is a standard normal distribution, as Equation 3.

$$\text{KL Divergence} = -\frac{1}{2}\sum(1 + \log \sigma^2 - \mu^2 - \sigma^2) \tag{3}$$

The total loss function is the sum of the reconstruction loss and the KL divergence. The **aim** is to balance accurate reconstruction and a well-regularized latent space. This balance ensures the model learns meaningful latent representations while avoiding overfitting the training data.

2.2.2. Vector Quantized VAE

The Vector Quantized Variational Autoencoder (VQ-VAE) is an addition to the regular VAE that adds discrete latent representations to give the latent space more structure and control. The standard VAE's latent space is continuous and Gaussian-distributed, but VQ-VAE's quantization is done with a codebook of discrete latent vectors. The architecture consists of an encoder, a decoder, and a quantization layer that connects them. In VQ-VAE, the encoder maps the input data x to a latent vector $z_e(x)$ in a continuous space. However, instead of directly passing this vector to the decoder, the latent vector is quantized to the nearest discrete value from a pre-defined codebook $e_i \epsilon R^D$, where i indexes the codebook entries and D is the dimensionality of each latent vector in the codebook. The quantization process can be expressed as Equation 4.

$$z_q(x) = argmin_{e_i} |z_e(x) - e_i|^2 \tag{4}$$

Here, the closest codebook vector e_i replaces the encoder's output $z_e(x)$, effectively discretizing the latent representation. The decoder then reconstructs the input using the quantized latent variable $z_q(x)$, aiming to minimize the reconstruction error. Reconstruction loss, vector quantization (VQ) loss, and commitment loss are the three main loss function components that drive the training of VQ-VAE. The reconstruction loss measures how accurately the decoder can reconstruct the original input data from the quantized latent vector. This is computed using Mean Squared Error (MSE) between the input x and its reconstruction $\hat{x}$, as (2). The VQ loss ensures that the codebook vectors move closer to the encoder's continuous latent output, keeping the quantization process effective. It measures the difference between the encoder output $z_e(x)$ and the closest codebook vector e_i, but the gradient does not flow through the encoder, as Equation 5. where "sg"(.) represents the stop-gradient operator, meaning the gradients do not backpropagate during the codebook update. This loss ensures that the codebook is updated to better match the encoder's latent space.

Here, the closest codebook vector e_i replaces the encoder's output $z_e(x)$, effectively discretizing the latent representation. The decoder then reconstructs the input using the quantized latent variable $z_q(x)$, aiming to minimize the reconstruction error. Reconstruction loss, vector quantization (VQ) loss, and commitment loss are the three main loss function components that drive the training of VQ-VAE. The reconstruction loss measures how accurately the decoder can reconstruct the original input data from the quantized latent vector. This is computed using Mean Squared Error (MSE) between the input x and its reconstruction $\hat{x}$ as (2). The VQ loss ensures that the codebook vectors move closer to the encoder's continuous latent output, keeping the quantization process effective. It measures the difference between the encoder output $z_e(x)$ and the closest codebook vector e_i, but the gradient does not flow through the encoder, as Equation 5. where "sg" (.) represents the stop-gradient operator, meaning the gradients do not backpropagate during the codebook update. This loss ensures that the codebook is updated to better match the encoder's latent space.

$$\text{VQ Loss} = |sg(z_e(x)) - e_i|^2 \tag{5}$$

The commitment loss encourages the encoder's output to commit to the discrete latent vectors in the codebook, preventing the encoder from straying too far from the quantized values. This term penalizes large discrepancies between the encoder output $z_e(x)$ and the codebook vector e_i, as Equation 6. The hyperparameter β controls the weight of the commitment loss, ensuring that the encoder's output adheres closely to the codebook while allowing for some flexibility in the latent space. The total loss in VQ-VAE combines these three terms as Equation 7. Each component is critical in ensuring the model learns a meaningful latent space while maintaining high-quality reconstructions. By introducing discrete latent representations, VQ-VAE offers better control over the generative process, making it especially useful in applications that benefit from structured data generation.

$$\text{VQ Loss} = \beta|z_e(x) - sg(e_i)|^2 \tag{6}$$

$$\text{Total Loss} = \text{Reconstruction Loss} + \text{VQ Loss } + \beta.\text{Commitment Loss} \tag{7}$$

2.3. Evaluation Metrics

When assessing the quality of synthetic data produced by models such as VAE and VQ-VAE, it is essential to employ quantitative metrics that evaluate the similarity between the synthetic data and the original data. Several well-established evaluation metrics

provide insight into different aspects of similarity between the datasets, focusing on the accuracy of individual data points and the overall distribution.

2.3.1. Mean Absolute Error (MAE)

The mean absolute error (MAE) is a simple yet effective metric for measuring the average absolute difference between the original data x and the synthetic data $\hat{x}$. It captures how far the predictions or generated data points deviate from the actual data without emphasizing large errors disproportionately, as Equation 8. Here, n represents the number of data points, x_i is the original value, and $\hat{x}_i$ is the corresponding synthetic value. MAE provides a straightforward measure of accuracy but treats all errors equally, making it useful when moderate errors are as significant as large ones.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{8}$$

2.3.2. Mean Squared Error (MSE) and Root Mean Squared Error (RMSE)

Mean Squared Error (MSE) builds on MAE by giving greater weight to larger errors, making it more sensitive to outliers or significant deviations between the original and synthetic data. It calculates the squared differences between corresponding data points and averages them as Equation 9. This approach ensures that large errors have a more pronounced effect on the overall metric, making it suitable for applications where larger deviations are undesirable. The root mean squared error (RMSE) is the square root of MSE, bringing the metric back to the same scale as the original data, making it easier to interpret Equation 10. This approach ensures that large errors have a more pronounced effect on the overall metric, making it suitable for applications where larger deviations are undesirable. The root mean squared error (RMSE) is the square root of MSE, bringing the metric back to the same scale as the original data, making it easier to interpret Equation 10.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \tag{9}$$

$$RMSE = \sqrt{MSE} \tag{10}$$

2.3.3. Coefficient of determination (R^2)

The coefficient of determination (R2) quantifies the extent to which the synthetic data accounts for the variance in the original dataset. The range is from 0 to 1, with a greater number signifying a superior alignment between the synthetic and original data, as stated in Equation 11. In this formula, $\bar{x}$ is the mean of the original data. A score of 1 indicates that the synthetic data perfectly replicates the variance of the original data. In contrast, a score of 0 means that the synthetic data provides no better fit than simply using the mean of the original data for every prediction. The R2 score is particularly useful when assessing how well the synthetic data captures the underlying structure and variability of the real data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2} \tag{11}$$

2.3.4. Kolmogorov-Smirnov (KS) Test

The Kolmogorov-Smirnov (KS) test is a statistical procedure that evaluates whether synthetic data originates from the same distribution as the original data. The two dataset empirical distribution functions (ECDF) are compared, and the biggest difference between them is calculated as Equation 12. where $F_n(x)$ and $F_{n'}(x)$ are the ECDFs of the original and synthetic data, respectively, and sup represents the supremum. The KS statistic $D_{n.n'}$ indicates the largest distance between the two distributions, with a smaller value suggesting a closer match. The associated p-value helps determine if the observed difference is statistically significant. The KS test is a powerful tool for assessing whether the synthetic data replicates the statistical properties of the original data, making it ideal for evaluating distributional similarity.

$$D_{n.n'} = sup_x|F_n(x) - F_{n'}(x)| \tag{12}$$

3. RESULT AND ANALYSIS

3.1. Loss Reduction

Figure 3 presents the loss curves for the standard VAE, while Figure 4 showcases the loss curves for the VQ-VAE. These figures provide insightful visualizations of each model's training dynamics and effectiveness through different aspects of their respective loss functions: total loss, reconstruction loss, KL divergence for VAE, and VQ loss and commitment loss for VQ-VAE. The Total Loss and Reconstruction Loss for the VAE model display a steep initial decrease, indicating rapid learning at the beginning of training. This is typical for VAEs, as they quickly adapt to the primary features of the dataset. The KL divergence, which measures the difference between the learned latent distribution and the prior distribution, also shows a sharp drop followed by a gradual decline, stabilizing as the model balances the latent space's structure with the reconstruction accuracy. This suggests that the VAE effectively learns a representation that captures the essential data characteristics while ensuring that the latent variables remain well-regularized.

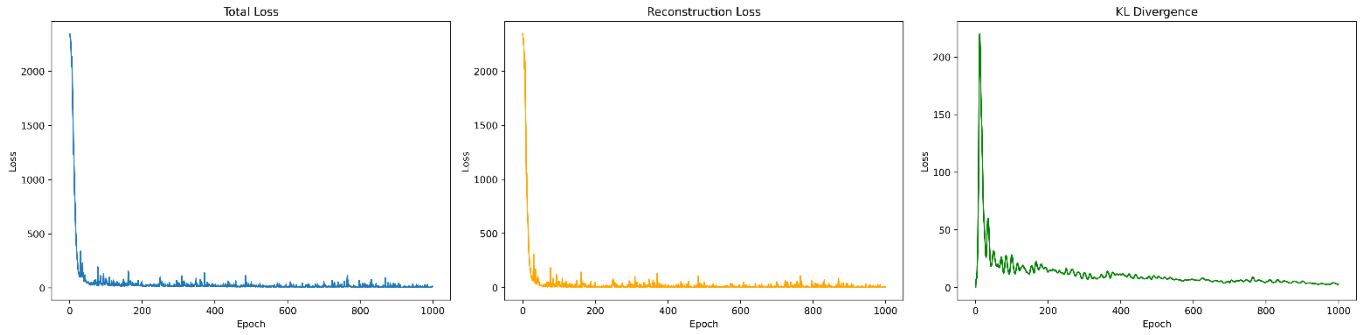


Figure 3. Loss reduction for VAE

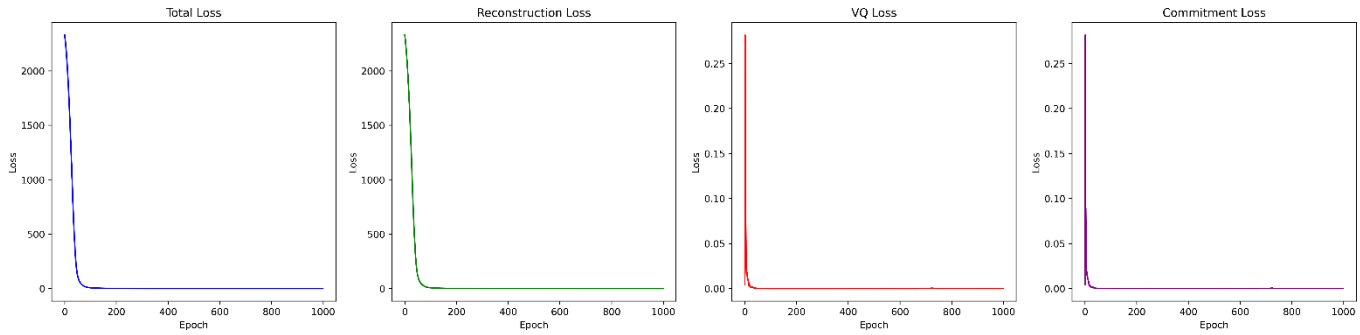


Figure 4. Loss reduction for VQ-VAE

In contrast, the VQ-VAE's Total Loss and Reconstruction Loss also decrease sharply, but the VQ Loss and Commitment Loss reveal distinct characteristics of the VQ-VAE's training process. The VQ Loss, which measures the error between the nearest codebook vector and the encoder output, decreases rapidly, indicating effective learning and adaptation of the codebook vectors. Commitment loss, penalizing the encoder's divergence from the selected codebook vector, also drops quickly, suggesting that the encoder efficiently commits to the appropriate vectors in the latent space. This rapid convergence in VQ and commitment losses shows that VQ-VAE quickly stabilizes, providing a robust and consistent latent representation.

Comparing both figures, the VQ-VAE exhibits a more segmented approach to loss reduction through its discrete latent variables, which likely contributes to its ability to generate more consistent and structured outputs. The VAE, with its continuous latent space, shows a smoother but potentially less controlled convergence in its KL divergence, pointing to a broader exploration of the latent

space but possibly at the cost of generating less precise synthetic data. The distinct behaviors observed in the VAE and VQ-VAE loss curves highlight their respective strengths and weaknesses: VAE's flexibility versus VQ-VAE's control and structure, impacting their suitability for different applications where either broader data generation capabilities or higher data fidelity is desired.

3.2. Synthetic Data Evaluation

The evaluation metrics in Table 1 for the VAE model show good performance across the open, high, low, and close features, with R2 values close to 0.98. This means that the model captures much of the original data's variation. However, the MAE, MSE, and RMSE are significantly higher for these features than those typically desired in financial modeling, pointing to discrepancies between the predicted and actual values. Notably, the KS statistic and the extremely low KS p-values indicate a significant difference between the original and synthetic data distributions, suggesting that the VAE struggles to match the data distribution, particularly at the tails, perfectly. The Volume feature shows much larger errors and a lower R2 value of 0.835, underscoring challenges in accurately modeling the volume, which may be due to its high volatility and irregular patterns.

Table 1. Performance Metrics of VAE

Feature	MAE	MSE	RMSE	R2	KS Statistic	KS P-Value
Open	8.963840e+01	9.181378e+03	9.581951e+01	0.979380	0.161451	1.163085e-43
High	9.198605e+01	9.712968e+03	9.855439e+01	0.978779	0.168551	1.448173e-47
Low	8.790611e+01	8.864689e+03	9.415248e+01	0.979504	0.164081	4.358578e-45
Close	8.923262e+01	9.151464e+03	9.566328e+01	0.979403	0.162503	3.147166e-44
Volume	2.087049e+07	4.990179e+14	2.233871e+07	0.835208	0.298712	4.738288e-150

In stark contrast, Table 2 shows that the VQ-VAE model achieves superior accuracy metrics across all features, with R^2 values nearing perfect scores of 0.999998 for price features and 0.999926 for volume. The MAE, MSE, and RMSE for open, high, low, and close are considerably lower. This demonstrates the VQ-VAE's enhanced capability to produce synthetic data that closely aligns with the original values. The very small KS statistics and perfect KS p-values for most features suggest no significant statistical difference between the original and synthetic data distributions, indicating an excellent match. Even though the Volume feature has big mistakes like the VAE model, it still has a higher R2 score and a slightly better KS statistic, which means it captures volume data variability better, but not perfectly.

Table 2. Performance Metrics of VQ-VAE

Feature	MAE	MSE	RMSE	R2	KS Statistic	KS P-Value
Open	0.099809	8.521067e-01	0.923096	0.999998	0.006048	1.000000
High	0.088027	7.124840e-01	0.844088	0.999998	0.005785	1.000000
Low	0.106512	8.912475e-01	0.944059	0.999998	0.006837	0.999992
Close	0.100404	8.609354e-01	0.927866	0.999998	0.005522	1.000000
Volume	82990.549040	2.233210e+11	472568.532477	0.999926	0.014462	0.821368

Comparing Tables 1 and 2, it is evident that the VQ-VAE model outperforms the standard VAE in almost all aspects. The major improvement seen with the VQ-VAE is attributed to its use of a discrete latent space, which allows for a more controlled and accurate data generation process, particularly effective in handling the complexities and nuances of financial time series data. The difference in performance is most pronounced in the lower MAE, MSE, and RMSE values achieved by the VQ-VAE, as well as the near-perfect R2 and KS p-values, which indicate not only close matches in value accuracy but also the overall distributional characteristics. This analysis highlights the suitability of VQ-VAE for applications requiring high fidelity in synthetic data generation, especially in fields like financial forecasting, where precise data replication is critical. VQ-VAE's ability to closely emulate the statistical properties of real data offers promising implications for its use in risk management, algorithmic trading, and economic modeling, where understanding and predicting complex market dynamics are crucial.

3.3. Visualization of Pattern Replications

The visualizations in Figure 5 and Figure 6 compare how each model reproduces the original time series data for stock features such as open, high, low, close, and volume. These figures help us understand how effectively the models capture the subtleties of stock market fluctuations. The VAE model demonstrates a commendable ability to follow the general trends and cyclical patterns

present in the stock data. The synthetic overlays on the original data for the open, high, low, and close features show that the VAE can approximate the major movements reasonably well. However, some deviations are noticeable during peaks and troughs, suggesting potential weaknesses in capturing extreme market behaviors. The volume graph reveals more pronounced disparities, with the synthetic data failing to accurately replicate the sharp spikes and volume surges. This discrepancy highlights the VAE's limitations in modeling highly volatile aspects of the data, which are critical in financial analyses.

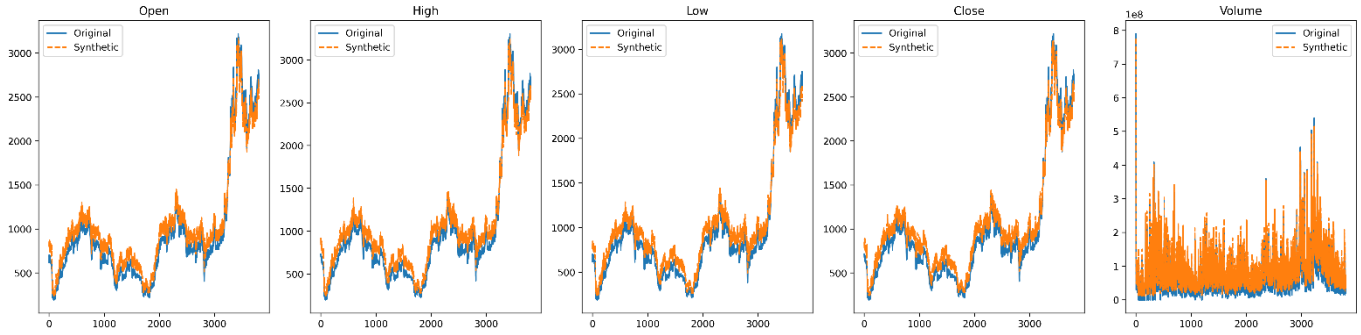


Figure 5. Pattern replication for VAE

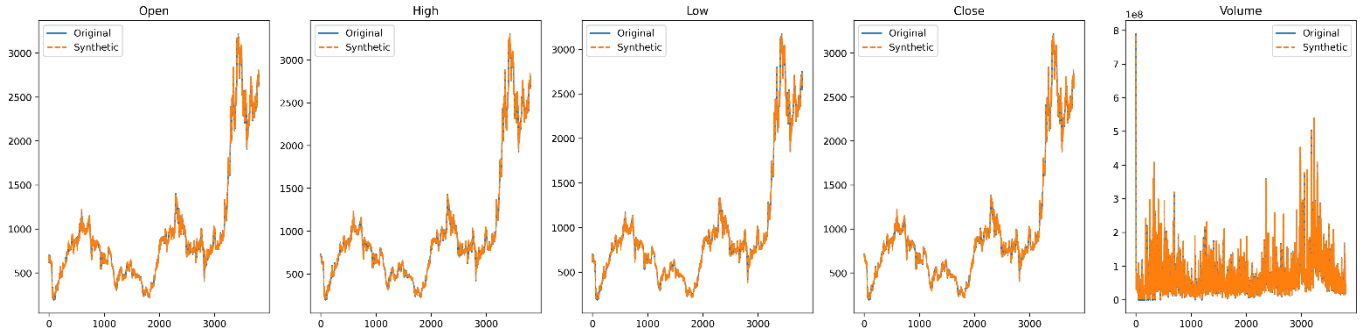


Figure 6. Pattern replication for VQ-VAE

In contrast, the VQ-VAE model aligns more closely with the original data across all features. The synthetic traces follow the general trends and match the peaks and valleys with greater fidelity than the VAE model. This improved performance is particularly evident in the low and close price charts, and trans the VQ-VAE synthetic data mimics the original data's contour more tightly. Though still challenging, the volume feature shows a better replication of volume changes. However, it still struggles with the highest spikes, a common difficulty in synthetic data generation, dealing with extreme values. Comparing Figures 5 and 6, it's evident that the VQ-VAE offers superior performance in terms of structural and temporal accuracy. The discrete latent space of the VQ-VAE likely contributes to this by enforcing a more organized and consistent representation, which is particularly advantageous in handling the complexities of financial time series data. While both models can capture the overarching patterns, the VQ-VAE's ability to more accurately model the nuances and extremities of the data suggests its greater utility for applications requiring high precision, such as predictive modeling and risk assessment in finance. This analysis underscores the importance of choosing the right model based on the specific requirements for fidelity and granularity in the synthetic data.

3.4. Visualization of Relationships and Distributions

Figures 7 and 8 showcase pair plot visualizations of synthetic and original data for the VAE and VQ-VAE models. These visualizations provide a detailed comparative analysis of how each model captures the relationships and distributions of various stock market features: open, high, low, close, and volume. The scatter plots for features like open, high, low, and close closely align along the diagonal in Figure 7, which shows that the VAE model generally demonstrates a good correlation between the original and synthetic data. The density plots show some differences, especially in representing the highest densities and the ends of the

distributions. This indicates that while the VAE can effectively estimate the average of the data, it has difficulty capturing the extreme values, which are important for risk management and predictive analytics. The volume scatter plot and density plot demonstrate noticeable differences from the original data, suggesting difficulties in accurately modeling volume changes, which can be highly volatile and influenced by external market factors.

Figure 7. Relationship and distribution for VAE

The VQ-VAE model visualizations in Figure 8 show a tighter alignment in the scatter and density plots across all features. The synthetic data generated by the VQ-VAE adheres more closely to the identity line in scatter plots and better matches the original data's distribution in density plots, including capturing the tails more effectively. This improved performance indicates the VQ-VAE's superior ability to structure and quantify the latent space to produce more accurate synthetic replicas of the original data. The VQ-VAE still shows some spread for the volume feature, but is noticeably better at clustering around the higher data points than the VAE model.

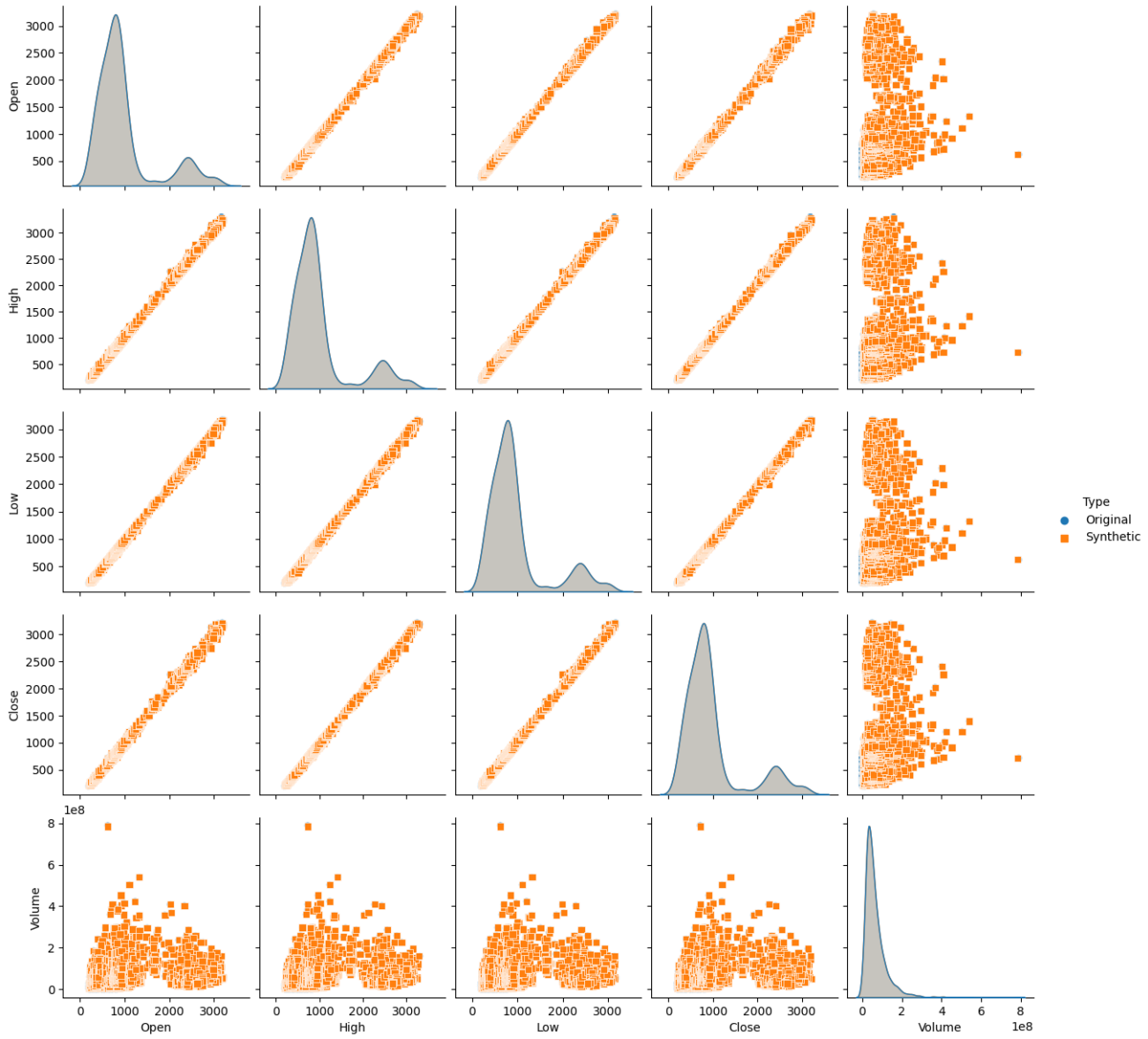


Figure 8. Relationship and distribution for VQ-VAE

Comparing Figures 7 and 8, the VQ-VAE consistently outperforms the VAE regarding data fidelity across all features. The discrete latent space of the VQ-VAE facilitates better control and replication of complex statistical properties, especially in capturing extremes and anomalies within the data, attributes crucial for accurate financial modeling. While flexible, the VAE's continuous latent space appears less capable of handling the nuances and spikes typical in financial time series data, as reflected in the broader dispersion and mismatch in the density profiles. This analysis underscores the VQ-VAE's potential in applications requiring high precision and reliability in synthetic data generation, such as in financial forecasting, risk assessment, and algorithmic trading, where even minor inaccuracies can lead to significantly different outcomes.

3.5. Comparison of statistical measures

We observe noticeable discrepancies in the statistical measures by comparing Table 3 (original data) with Table 4 (synthetic data from the VAE model). The mean values for open, high, low, and close in the synthetic data are lower than those of the original,

indicating that the VAE model underestimates these metrics. The standard deviation in the synthetic data is higher across all features except volume, suggesting that the VAE-generated data exhibits more variability than the original data, which could indicate less stability in the model's performance. The minimum values for open, high, low, and close are similar, showing that the VAE can replicate the lowest values reasonably well. However, the significantly lower 25th percentile and median values across these features demonstrate a skew in the distribution of the synthetic data towards lower values. The volume in the synthetic data shows a substantial deviation, especially at the minimum, indicating that the VAE struggles to capture low-volume trading days accurately.

Table 3. Statistical Measurement of Original Data

	Open	High	Low	Close	Volume
count	3803.000000	3803.000000	3803.000000	3803.000000	3.803000e+03
mean	970.524323	986.688667	953.595057	969.293715	6.480496e+07
std	667.362514	676.630856	657.735744	666.656839	5.503598e+07
min	194.000000	198.000000	190.000000	194.000000	3.000000e+00
25%	567.000000	579.000000	556.500000	565.000000	3.012070e+07
50%	792.000000	804.000000	779.000000	790.000000	4.977740e+07
75%	999.000000	1011.000000	977.500000	994.000000	8.149770e+07
max	3220.000000	3305.000000	3173.000000	3220.000000	7.897860e+08

Table 4. Statistical Measurement of Synthetic Data Generated by VAE

	Open	High	Low	Close	Volume
count	3803.000000	3803.000000	3803.000000	3803.000000	3803.0
mean	902.278198	915.281067	885.929138	900.289856	44017412.0
std	723.759094	733.523376	712.361084	722.080322	52841740.0
min	198.399506	203.209579	194.313629	197.170486	893770.5
25%	450.684769	459.612793	440.914520	449.630264	13947248.5
50%	689.237366	698.285950	674.004211	686.924927	27158270.0
75%	919.798492	931.988922	905.348541	923.163452	52520946.0
max	3216.591309	3297.606201	3168.208496	3216.165039	787979072.0

Table 5. Statistical Measurement of Synthetic Data Generated by VQ-VAE

	Open	High	Low	Close	Volume
count	3803.000000	3803.000000	3803.000000	3803.000000	3.803000e+03
mean	970.572205	986.725220	953.641785	969.342468	6.484290e+07
std	667.294922	676.572266	657.650696	666.581909	5.497283e+07
min	208.552155	210.586380	202.853790	207.824173	3.331917e+06
25%	566.998535	579.000641	556.500214	565.000397	3.010851e+07
50%	792.000366	803.998474	778.998169	790.000122	4.977641e+07
75%	998.998505	1010.998749	977.499420	994.000214	8.149736e+07
max	3203.241943	3288.859375	3157.541260	3204.150146	7.862252e+08

The findings of this research are that the synthetic data generated by the VQ-VAE model (as shown in Table 5) more accurately reflects the statistical characteristics of the original dataset (Table 3) compared to the synthetic data generated by the standard VAE (Table 4). The mean and standard deviation values across the open, high, low, and close features in VQ-VAE are nearly identical to the original, indicating high fidelity in replicating central tendencies and volatility. Furthermore, the 25th, 50th (median), and 75th percentile values align closely with the actual data, confirming that VQ-VAE effectively preserves the distributional properties across the entire value range. While some challenges persist in reproducing extreme minimum values for volume, the VQ-VAE still demonstrates improved replication accuracy over the VAE, particularly in capturing higher volume observations, which are critical for modeling market behavior under peak activity.

The results of this research **are in line** with previous studies, such as research [19], who found that discrete latent space representations in generative models enhance the preservation of data structure in biomedical informatics. Similarly, research [24] demonstrated that using vector quantization in latent space increases consistency and reconstruction accuracy in speech synthesis, which aligns with our finding that VQ-VAE offers more structured and faithful replication of financial time series data. Compared to VAE, which tends to exhibit broader variability and underestimation of key metrics, our results show that VQ-VAE provides a more

robust mechanism for data generation under the constraints of financial modeling. These findings are also **supported** by [28], who emphasize that using discrete encoding strategies helps preserve statistical coherence and distributional fidelity in synthetic financial time series data. Therefore, the superiority of VQ-VAE in our experiment confirms these prior insights and extends their validity to Indonesian financial data, a previously underexplored context.

3.6. Strengths and weaknesses

The VQ-VAE introduces significant advancements in the control and structure of latent space representation, addressing some of the limitations inherent in traditional autoencoders [31]. By utilizing a discrete latent space, the VQ-VAE ensures that the variability in generated synthetic data is well-regulated and more predictable, enhancing the reproducibility and consistency of the outputs [32]. This structured approach to data generation is facilitated by the model's use of a fixed codebook of latent vectors, which the encoder maps input data onto [19]. This quantization process stabilizes the learning environment and allows for precise reconstruction, making the VQ-VAE particularly effective in domains where maintaining the integrity of data features is crucial [25]. Such a methodologically rigid structure tends to produce synthetic data that adheres closely to the statistical properties of the training set, thereby improving the reliability of simulations and analyses derived from the synthetic data [33]. However, the strengths of the VQ-VAE come with trade-offs, primarily related to computational efficiency. Mapping high-dimensional data to discrete latent vectors and the subsequent quantization introduces additional complexity and computational overhead [34]. As a result, the VQ-VAE often requires more processing power and longer training times compared to models with continuous latent spaces [31]. This increased computational demand can be a significant drawback in scenarios where training time and resource efficiency are critical [24]. Furthermore, the rigidity of the quantized latent space, while beneficial for consistency, may limit the model's flexibility in capturing the more nuanced or subtle variations within the data that do not align neatly with the predefined vectors in the codebook [25].

Despite these challenges, the VQ-VAE holds considerable promise for various applications, especially within financial forecasting and other time series analyses. Its ability to generate structured, consistent data makes it well-suited for modeling complex financial markets, where the accuracy and reliability of predictive analytics can significantly impact decision-making and strategic planning [24]. In time series forecasting, the VQ-VAE's discrete representation can effectively model seasonal trends, cyclical patterns, and irregular fluctuations, providing a robust foundation for forecasting models [25]. Moreover, the VQ-VAE could be instrumental in risk management and algorithmic trading, where the generation of synthetic but realistic market scenarios can help stress test and evaluate investment strategies under various simulated conditions. Thus, the VQ-VAE's unique capabilities make it an attractive option for financial analysts and data scientists seeking precise and dependable modeling tools[33].

4. CONCLUSION

In conclusion, the comparative analysis between the standard VAE and the VQ-VAE on the ADRO stock dataset has highlighted significant advantages of the VQ-VAE in generating synthetic data. The key findings suggest that VQ-VAE, with its discrete latent space, consistently outperforms the standard VAE in producing higher-quality and more accurate synthetic representations of the original data. This enhanced performance is attributed to the structured and controlled environment provided by the discrete latent space, which facilitates more precise and reliable data generation. Such attributes make the VQ-VAE particularly valuable in scenarios where the fidelity of data reproduction is paramount, as it ensures that the synthetic data retains the original dataset's critical statistical properties and structures without introducing unwanted variability. The implications of these findings for the field of finance, particularly in generative models, are substantial. The VQ-VAE's ability to generate structured and controlled synthetic data positions it as a superior tool for financial forecasting and complex financial analyses where accuracy and reliability are crucial. Looking ahead, there are promising avenues for further enhancing the capabilities of VQ-VAE. Integrating attention mechanisms or generative adversarial networks (GANs) with VQ-VAE could improve the adaptability and sensitivity of the model to finer nuances in data, leading to even higher-quality synthetic outputs. Moreover, exploring the application of VQ-VAE to other datasets in domains like insurance or fintech could reveal more potential benefits and versatility in synthetic data generation across various fields. This expansion could lead to broader adoption and innovation in how synthetic data is used to solve complex problems in industries reliant on large and dynamic datasets.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AUTHOR CONTIBUTION

FUNDING STATEMENT

COMPETING INTEREST

REFERENCES

[1] A. O. Aseeri, "Effective short-term forecasts of Saudi stock price trends using technical indicators and large-scale multivariate time series," *PeerJ Computer Science*, vol. 9, p. e1205, Jan. 2023, https://doi.org/10.7717/peerj-cs.1205.

[2] K. Alkhatib, H. Khazaleh, H. A. Alkhazaleh, A. R. Alsoud, and L. Abualigah, "A New Stock Price Forecasting Method Using Active Deep Learning Approach," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 2, p. 96, Jun. 2022, https://doi.org/10.3390/joitmc8020096.

[3] Y.-C. Chen and W.-C. Huang, "Constructing a stock-price forecast CNN model with gold and crude oil indicators," *Applied Soft Computing*, vol. 112, p. 107760, Nov. 2021, https://doi.org/10.1016/j.asoc.2021.107760.

[4] A. U. Haq, A. Zeb, Z. Lei, and D. Zhang, "Forecasting daily stock trend using multi-filter feature selection and deep learning," *Expert Systems with Applications*, vol. 168, p. 114444, Apr. 2021, https://doi.org/10.1016/j.eswa.2020.114444.

[5] Z. Zhang and M. Wu, "Predicting Real-Time Locational Marginal Prices: A GAN-Based Approach," *IEEE Transactions on Power Systems*, vol. 37, no. 2, pp. 1286–1296, Mar. 2022, https://doi.org/10.1109/TPWRS.2021.3106263.

[6] L. B. Iantovics and C. Enăchescu, "Method for Data Quality Assessment of Synthetic Industrial Data," *Sensors*, vol. 22, no. 4, p. 1608, Feb. 2022, https://doi.org/10.3390/s22041608.

[7] S. Tuarob, P. Wettayakorn, P. Phetchai, S. Traivijitkhun, S. Lim, T. Noraset, and T. Thaipisutikul, "DAViS: A unified solution for data collection, analyzation, and visualization in real-time stock market prediction," *Financial Innovation*, vol. 7, no. 1, p. 56, Dec. 2021, https://doi.org/10.1186/s40854-021-00269-7.

[8] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," *Journal of Big Data*, vol. 7, no. 1, p. 66, Dec. 2020, https://doi.org/10.1186/s40537-020-00333-6.

[9] X. Hou, K. Wang, C. Zhong, and Z. Wei, "ST-Trader: A Spatial-Temporal Deep Neural Network for Modeling Stock Market Movement," *IEEE/CAA Journal of Automatica Sinica*, vol. 8, no. 5, pp. 1015–1024, May 2021, https://doi.org/10.1109/JAS.2021.1003976.

[10] A. H. Bukhari, M. A. Z. Raja, M. Sulaiman, S. Islam, M. Shoaib, and P. Kumam, "Fractional Neuro-Sequential ARFIMA-LSTM for Financial Market Forecasting," *IEEE Access*, vol. 8, pp. 71 326–71 338, 2020, https://doi.org/10.1109/ACCESS.2020.2985763.

[11] H. Gunduz, "An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination," *Financial Innovation*, vol. 7, no. 1, p. 28, Dec. 2021, https://doi.org/10.1186/s40854-021-00243-3.

[12] H. Li, Y. Cui, S. Wang, J. Liu, J. Qin, and Y. Yang, "Multivariate Financial Time-Series Prediction With Certified Robustness," *IEEE Access*, vol. 8, pp. 109 133–109 143, 2020, https://doi.org/10.1109/ACCESS.2020.3001287.

[13] D. Panfilo, A. Boudewijn, S. Saccani, A. Coser, B. Svara, C. R. Chauvenet, C. A. Mami, and E. Medvet, "A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data," *IEEE Access*, vol. 11, pp. 63 306–63 323, 2023, https://doi.org/10.1109/ACCESS.2023.3288336.

[14] Y. Jin, R. McDaniel, N. J. Tatro, M. J. Catanzaro, A. D. Smith, P. Bendich, M. B. Dwyer, and P. T. Fletcher, "Implications of data topology for deep generative models," *Frontiers in Computer Science*, vol. 6, p. 1260604, Aug. 2024, https://doi.org/10.3389/fcomp.2024.1260604.

[15] Y. L. Chow, S. Singh, A. E. Carpenter, and G. P. Way, "Predicting drug polypharmacology from cell morphology readouts using variational autoencoder latent space arithmetic," *PLOS Computational Biology*, vol. 18, no. 2, p. e1009888, Feb. 2022, https://doi.org/10.1371/journal.pcbi.1009888.

[16] B. Hernandez, O. Stiff, D. K. Ming, C. Ho Quang, V. Nguyen Lam, T. Nguyen Minh, C. Nguyen Van Vinh, N. Nguyen Minh, H. Nguyen Quang, L. Phung Khanh, T. Dong Thi Hoai, T. Dinh The, T. Huynh Trung, B. Wills, C. P. Simmons, A. H. Holmes, S. Yacoub, P. Georgiou, and on behalf of the Vietnam ICU Translational Applications Laboratory (VITAL) investigators, "Learning meaningful latent space representations for patient risk stratification: Model development and validation for dengue and other acute febrile illness," *Frontiers in Digital Health*, vol. 5, p. 1057467, Feb. 2023, https://doi.org/10.3389/fdgth.2023.1057467.

[17] M. Nabipour, P. Nayyeri, H. Jabani, S. S., and A. Mosavi, "Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis," *IEEE Access*, vol. 8, pp. 150 199–150 212, 2020, https://doi.org/10.1109/ACCESS.2020.3015966.

[18] J. Wu, K. Plataniotis, L. Liu, E. Amjadian, and Y. Lawryshyn, "Interpretation for Variational Autoencoder Used to Generate Financial Synthetic Tabular Data," *Algorithms*, vol. 16, no. 2, p. 121, Feb. 2023, https://doi.org/10.3390/a16020121.

[19] R. Wei and A. Mahmood, "Recent Advances in Variational Autoencoders With Representation Learning for Biomedical Informatics: A Survey," *IEEE Access*, vol. 9, pp. 4939–4956, 2021, https://doi.org/10.1109/ACCESS.2020.3048309.

[20] Z. Feng, M. Daković, H. Ji, X. Zhou, M. Zhu, X. Cui, and L. Stanković, "Interpretation of Latent Codes in InfoGAN with SAR Images," *Remote Sensing*, vol. 15, no. 5, p. 1254, Feb. 2023, https://doi.org/10.3390/rs15051254.

[21] S. Saha, F. Bovolo, and L. Bruzzone, "Building Change Detection in VHR SAR Images via Unsupervised Deep Transcoding," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021, https://doi.org/10.1109/TGRS.2020.3000296.

[22] M. Elbattah, C. Loughnane, J.-L. Guérin, R. Carette, F. Cilia, and G. Dequen, "Variational Autoencoder for Image-Based Augmentation of Eye-Tracking Data," *Journal of Imaging*, vol. 7, no. 5, p. 83, May 2021, https://doi.org/10.3390/jimaging7050083.

[23] X. Jiang, X. Peng, H. Xue, Y. Zhang, and Y. Lu, "Latent-Domain Predictive Neural Speech Coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2111–2123, 2023, https://doi.org/10.1109/TASLP.2023.3277693.

[24] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He, S. Zhao, T. Qin, F. Soong, and T.-Y. Liu, "*NaturalSpeech* : End-to-End Text-to-Speech Synthesis With Human-Level Quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4234–4245, Jun. 2024, https://doi.org/10.1109/TPAMI.2024.3356232.

[25] A. Asperti, L. Bugo, and D. Filippini, "Enhancing Variational Generation Through Self-Decomposition," *IEEE Access*, vol. 10, pp. 67 510–67 520, 2022, https://doi.org/10.1109/ACCESS.2022.3185654.

[26] L. Li, J. Yan, H. Wang, and Y. Jin, "Anomaly Detection of Time Series With Smoothness-Inducing Sequential Variational Auto-Encoder," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1177–1191, Mar. 2021, https://doi.org/10.1109/TNNLS.2020.2980749.

[27] Y. Liu, W. Xie, Y. Li, Z. Li, and Q. Du, "Dual-Frequency Autoencoder for Anomaly Detection in Transformed Hyperspectral Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022, https://doi.org/10.1109/TGRS.2022.3152263.

[28] M. Dogariu, L.-D. Ştefan, B. A. Boteanu, C. Lamba, B. Kim, and B. Ionescu, "Generation of Realistic Synthetic Financial Time-series," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 4, pp. 1–27, Nov. 2022, https://doi.org/10.1145/3501305.

[29] M. Diqi, E. Utami, K. Kusrini, and F. W. Wibowo, "Indonesian Stocks," https://doi.org/10.34740/KAGGLE/DSV/8357240.

[30] J. I. Monroe and V. K. Shen, "Systematic control of collective variables learned from variational autoencoders," *The Journal of Chemical Physics*, vol. 157, no. 9, p. 094116, Sep. 2022, https://doi.org/10.1063/5.0105120.

[31] H. Guo, F. Xie, X. Wu, F. K. Soong, and H. Meng, "MSMC-TTS: Multi-Stage Multi-Codebook VQ-VAE Based Neural TTS," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1811–1824, 2023, https://doi.org/10.1109/TASLP.2023.3272470.

[32] K. El Emam, L. Mosquera, X. Fang, and A. El-Hussuna, "An evaluation of the replicability of analyses using synthetic health data," *Scientific Reports*, vol. 14, no. 1, p. 6978, Mar. 2024, https://doi.org/10.1038/s41598-024-57207-7.

[33] X. Q. Chen, L. Zhang, and T. J. Cui, "Intelligent autoencoder for space-time-coding digital metasurfaces," *Applied Physics Letters*, vol. 122, no. 16, p. 161702, Apr. 2023, https://doi.org/10.1063/5.0132635.

[34] R. Wei, C. Garcia, A. El-Sayed, V. Peterson, and A. Mahmood, "Variations in Variational Autoencoders - A Comparative Evaluation," *IEEE Access*, vol. 8, pp. 153 651–153 670, 2020, https://doi.org/10.1109/ACCESS.2020.3018151.