IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

Deep Learning Factor Investing in the Indonesian Stock Market

Abstrak

Model faktor linear tradisional seringkali gagal menangkap dinamika pasar saham negara berkembang yang kompleks dan non-linear. Penelitian ini merancang dan memvalidasi metodologi deep learning baru menggunakan matriks Recurrence Plot (RP) dengan β-VAE untuk menemukan faktor investasi non-linear dalam konteks Indonesia. Kami menunjukkan bahwa kerangka kerja ini secara sistematis merupakan "pabrik faktor" yang unggul dibandingkan dengan baseline linear RP dengan PCA, menemukan dua kali lebih banyak faktor berkualitas tinggi (Sharpe > 0.3) dan menghasilkan alpha rata-rata 7 kali lipat lebih tinggi. Temuan kunci adalah kemampuan model untuk memisahkan sinyal prediktif frekuensi tinggi (diidentifikasi oleh SHAP) dari tren profitabel frekuensi rendah yang lebih berharga (divalidasi melalui backtesting). Faktor juara dari proses ini menghasilkan alpha tahunan yang kuat sebesar 6.65% dengan max drawdown minimal -7.73% dari 2018 hingga 2025. Studi ini menyimpulkan bahwa pendekatan RP dengan β-VAE adalah kerangka kerja yang kuat dan tangguh untuk menemukan sumber return non-linear yang lebih aman dan tidak dapat dijelaskan oleh model konvensional.

Kata kunci— *Factor Investing, Deep Learning, Variational Autoencoder, Asset Pricing, Pasar Modal Indonesia*

Abstract

Traditional linear factor models often fail to capture the complex, non-linear dynamics of emerging stock markets. This research designs and validates a novel Recurrence Plot (RP) matrices with β-VAE deep learning methodology to discover non-linear investment factors within the Indonesian context. We demonstrate that this framework is a systematically superior "factor factory" compared to a linear RP with PCA baseline, discovering twice as many high-quality factors (Sharpe > 0.3) and generating 7-fold more alpha on average. A key finding is the model's ability to disentangle high-frequency predictive signals (identified by SHAP) from more valuable, low-frequency profitable trends (validated by backtesting). The champion factor from this process yields a robust annualized alpha of 6.65% with a minimal max drawdown of -7.73% from 2018 to 2025. This study concludes that the RP -> β -VAE approach is a robust and resilient framework for discovering safer, non-linear sources of return unexplained by conventional models.

Keywords—*Factor Investing, Deep Learning, Variational Autoencoder, Asset Pricing, Indonesian Capital Market*

1. INTRODUCTION

The field of factor investing is undergoing a significant transformation, driven by two critical challenges. First, the proliferation of hundreds of documented factors has led to what is known as the "factor zoo," where many signals fail to demonstrate robustness out-of-sample, raising questions about data mining and statistical validity. Second, a growing body of evidence

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

from comprehensive studies has established that stock return dynamics are fundamentally non-linear, a characteristic that traditional linear models are ill-equipped to capture [1].This limitation is particularly pronounced in emerging markets like Indonesia, whose unique structural properties and volatility regimes demand more sophisticated, data-driven approaches [2].

While machine learning has emerged as a powerful tool to address non-linearity, its application in asset pricing has predominantly been in a supervised context for direct return forecasting. This "black box" approach, while often accurate, provides limited insight into the underlying economic drivers of returns, hindering its utility for true factor discovery [1]. A significant methodological gap, therefore, exists in the application of unsupervised deep generative models, which seek not merely to predict but to learn the underlying data-generating process of market returns.

The core limitation of existing unsupervised methods, such as Principal Component Analysis (PCA), lies in their inability to adequately process the rich, dynamic information embedded in financial time-series. These models typically operate on one-dimensional return data, failing to capture complex temporal dependencies, state transitions, and volatility clustering. Furthermore, they often produce latent factors that are entangled statistically correlated and difficult to interpret as distinct economic signals [3], [4], thus failing to provide the clean, independent risk exposures desired for modern portfolio construction.

To address this challenge, this research designs and validates a novel quantitative pipeline that synergizes advanced feature engineering with a disentangled deep generative model. We hypothesize that the limitations of existing methods can be overcome by first transforming one-dimensional return series into a richer, two-dimensional representation: a Recurrence Plot (RP) matrix. An RP functions as a "behavioral fingerprint" of an asset's dynamics [5], converting temporal patterns into a structured image that a convolutional neural network can effectively analyze. This richer representation is then processed by a β-Variational Autoencoder (β-VAE), an architecture explicitly designed to learn a disentangled latent space and a proven method for feature extraction in financial time series [6].

The primary contribution of this research is three-fold. First, we introduce a novel and robust methodology that integrates Recurrence Plots with β-VAEs for non-linear factor discovery. Second, we prove its systematic superiority over a linear PCA baseline. Our ablation study demonstrates that our method is a more reliable "factor factory," discovering twice as many high-quality factors and producing 7-fold more alpha on average while yielding safer, more resilient signals with lower drawdowns. Third, we uncover a crucial distinction between short-term predictive power (identified by SHAP) [7] and long-term economic profit (validated by backtesting), leveraging this to isolate a persistent, non-linear alpha signal of 6.65% and confirming a new, exploitable source of low-frequency return in the Indonesian capital market.

2. METHODS

This research employs a multi-stage quantitative pipeline to discover, validate, and backtest non-linear investment factors, as illustrated in Figure 1.

Figure 1 Research Methodology

The methodology is designed to be end-to-end, beginning from raw price data and culminating in a comprehensive performance evaluation.

2.1 Data Preparation

The pipeline begins with data preparation. To avoid survivorship bias and ensure a tradable universe, a critical step in machine learning factor discovery (Han, He, & Zhou, 2023).

We construct a dynamic monthly universe of stocks listed on the Indonesia Stock Exchange (IDX) from January 2018 to September 2025. The universe is reconstructed monthly with the following inclusion criteria: 1. Stocks must meet a market capitalization screen of IDR 1 Trillion (1,000,000,000,000) or greater at the time of reconstruction. This ensures the universe consists of sufficiently large and liquid companies. 2. To handle IPOs (the "ragged start" issue), a stock must have at least 252 trading days of price history to be included in the universe. This ensures all data inputs are sufficiently long.

This robust filtering process results in a dynamic universe of eligible stocks for each month. From this clean panel data, a matrix of daily log returns was computed. External macroeconomic data, including the BI 7-Day Repo Rate and commodity prices, were also collected for the same period.

2.2 Feature Engineering

The second stage is feature engineering, where one-dimensional log return series are transformed into two-dimensional Recurrence Plot (RP) matrices. This transformation is designed to capture complex, non-linear temporal dependencies and market regimes, converting the "behavioral fingerprint" of an asset's dynamics into a structured image format suitable for convolutional analysis [8], [9]. The RP is a matrix R where each element $R_{i,j}$ is defined in equation (1):

$$R_{i,j} = \Theta\left(\epsilon - \left|\vec{x_i} - \vec{x_j}\right|\right) \tag{1}$$

Where $\vec{x_i}$ and $\vec{x_j}$ are vectors in the 15-dimensional phase space, $||\cdot||$ is a norm (e.g., Euclidean distance), ϵ is a distance threshold, and $\Theta(\cdot)$ is the Heaviside step function. In essence, the matrix element is 1 (plotted as a dot) if two points in the time-series' trajectory are "close" and 0 otherwise. For each stock, a sequence of 64x64 RP matrices is generated using a rolling window of 252 trading days (with a step size of 21 days) on a 15-dimensional phase space embedding of the returns.

2.3 Factor Discovery (Models)

Following the feature engineering stage, the core factor discovery is performed using two different unsupervised models. This stage aims to learn a compressed, meaningful latent representation from the high-dimensional RP matrices.

2. 3.1 Our Proposed Model (β-VAE)

The primary model is a β-Variational Autoencoder (β -VAE), an architecture specifically designed to learn a disentangled latent representation for financial factor analysis (Poh & Wang, 2022; Wang, Wang, & Li, 2023).

The model employs a deep convolutional structure. The encoder consists of an initial 2D convolution followed by a series of four residual down-sampling blocks (ResDown) and one residual block (ResBlock) to progressively reduce dimensionality, culminating in two separate 1x1 convolutions to output the parameters (μ and $\log\sigma^2$) of the latent distribution. The decoder mirrors this structure using residual up-sampling blocks (ResUp) and a residual block (ResBlock) to reconstruct the RP matrix from the sampled latent vector. Both encoder and decoder utilize the Exponential Linear Unit (ELU) activation function and Batch Normalization.

Unlike a standard autoencoder, the β-VAE is trained to optimize a modified Evidence Lower Bound (ELBO), explicitly balancing reconstruction accuracy and latent space regularization, as shown in equation (2):

$$\mathcal{L} = E[\log p\,(x|z)] - \beta D_{KL}\big(q(z|x)|p(z)\big) \tag{2}$$

The first term represents the Reconstruction Loss, measured using Mean Squared Error (MSE) between the input RP x and the reconstructed RP $\hat{x}$. The second term is the KL-Divergence between the learned latent distribution $\mathrm{q(z|x)}$ and a standard normal prior $p(z)$, weighted by the hyperparameter $\beta > 1$ to enforce a disentangled latent space z. For this study, we train the model to learn a 10-dimensional latent representation (i.e., 10 factors).

The model is trained end-to-end in an unsupervised manner for 50 epochs. We use the Adam optimizer with an initial learning rate (lr) of 1e-4 and weight decay of 1e-5. A Cosine Annealing learning rate scheduler is employed to adjust the learning rate during training.

2. 3.2 Baseline Model (PCA)

To validate the choice of the complex β-VAE, we also construct a linear baseline model using Principal Component Analysis (PCA). This model is fed the identical 64x64 Recurrence Plot matrices. We then extract the top 10 principal components (i.e., 10 factors) to ensure a direct comparison against the 10 latent factors from the β-VAE.

2. 3.3 Hyperparameter Tuning and Out-of-Sample Application

Critically, to prevent data leakage, model hyperparameters specific to the β -VAE (such as the β value) were tuned using a separate validation set. This validation set was constructed by randomly holding out 20% of the stocks from the training process. The final β -VAE model configuration was selected based on its optimal trade-off between reconstruction loss and a disentanglement metric on this validation set, not on backtest performance. Both the chosen β -VAE model and the PCA model were then applied once to generate factors across the full out-of-sample period (using all stocks) before proceeding to the validation and evaluation stages.

2.4 Predictive Validation

The fourth stage is predictive validation, a two-step process. First, a correlation analysis provides an initial economic interpretation. Second, the factors' marginal predictive contribution is rigorously quantified. A LightGBM model is trained to predict next-month's neutralized returns. The marginal contribution of each factor is then precisely quantified using TreeSHAP [10].

2.5 Economic Evaluation

Finally, the economic significance of the factors from both the β-VAE and PCA models is evaluated through a comprehensive long-short quintile portfolio backtest (Fama & French, 1993), a methodology validated for use in emerging markets (Jacobs & Müller, 2020). The performance of these strategies is measured using a robust set of standard industry metrics: Annualized Sharpe Ratio, Annualized Alpha, Annualized Volatility, and Max Drawdown. Furthermore, the robustness of the discovery process itself is evaluated by comparing the 'hit rate' (number of positive-alpha factors) and the average alpha across all discovered factors, aligning with best practices for robustness analysis in deep learning asset pricing [11].

3. RESULTS AND DISCUSSION

This section presents the empirical findings from our multi-stage quantitative pipeline, detailing model selection, economic interpretation, SHAP validation, and backtesting.

The core of our methodology is the β -VAE, which was trained to process a large-scale dataset of over 26,000 Recurrence Plot (RP) matrices (one for each monthly window across our dynamic stock domain) to learn a compressed, 10-dimensional latent representation. As described in our Methods, to prevent data leakage, the β hyperparameter was tuned on a separate validation set by optimizing for a trade-off between reconstruction loss and a disentanglement metric. A β of 3.0 was selected as it provided the most stable and well-disentangled latent space. This final, pre-selected model was then run once across the entire period to generate the panel dataset of 10 latent factor time series, spanning from late 2018 to September 2025.

To provide an initial economic interpretation, we correlated the market-wide average of each factor against the BI 7-Day Repo Rate and IDR-denominated coal and palm oil prices. The results are presented in Figure 2.

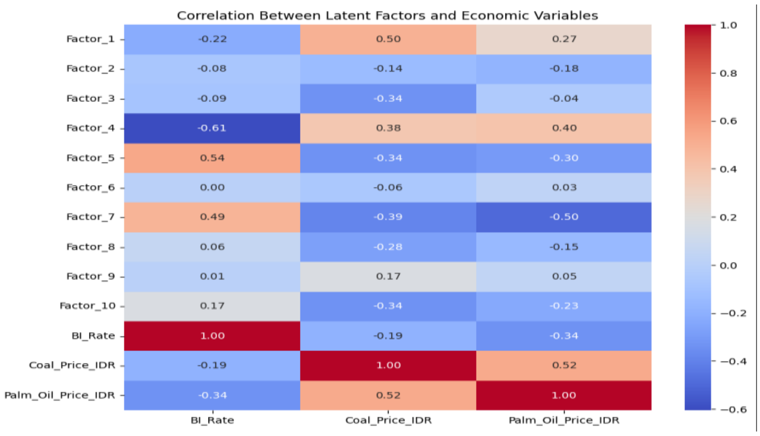

Figure 2 Correlation of latent factors with economic variables

Figure 2 shows that the key finding from this analysis is that the discovered factors are not simple linear proxies for the macroeconomic variables. For instance, Factor_4 exhibits a strong negative correlation of -0.61 with the BI_Rate, clearly identifying it as a potent "Monetary Policy" or "Interest Rate Sensitivity" factor. In contrast, Factor_1 shows the strongest positive relationship with Coal_Price_IDR (0.50), suggesting it captures a specific "Coal/Energy Sector" theme. Meanwhile, other factors like Factor_5 and Factor_7 display strong inverse relationships with the BI Rate and commodity prices, respectively, indicating they represent distinct aspects of these economic forces. This lack of a simple one-to-one mapping across all factors, combined with the emergence of these specialized signals, is a strong indication that the β-VAE has successfully decomposed complex market dynamics into more interpretable, underlying drivers.

To dig deeper into the nature of the economic correlations we found in Figure 2, we plotted the time series for the most correlated factors, Factor_1 (linked to Coal). The plot for Factor_1 is shown in Figure 3, as it perfectly illustrates the non-linear dynamics our model captured.

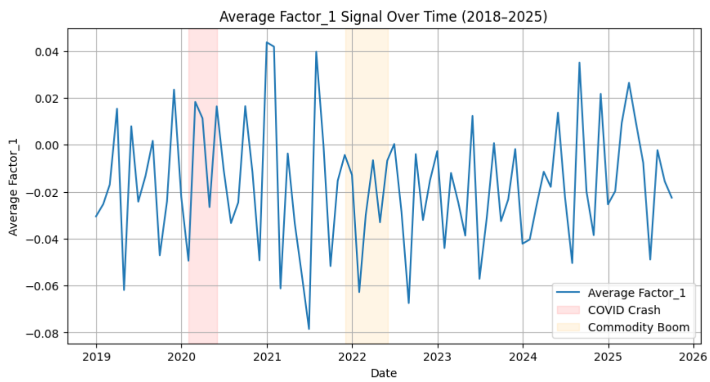

Figure 3 Market-Wide Average of Factor_1 Signal (2018 (late)-2025)

The plot shows in Figure 3 reveals that the +0.50 correlation is not a simple linear relationship. Instead of just trending with the 2022 commodity boom, the factor exhibits extreme volatility during the period, suggesting it captures the regime's uncertainty rather than just its price trend.

Furthermore, the factor's behavior around the 2020 COVID crash is telling. It plummets to a significant trough during the crash, followed by its largest-ever spike in the subsequent recovery phase (late 2020/early 2021). This demonstrates that the β -VAE has successfully isolated complex, non-linear signals that capture regime shifts (like crash-to-recovery) and volatility dynamics, that a simple linear correlation value fails to express. We will now test the predictive and economic value of these complex factors.

To rigorously quantify the predictive power of our discovered factors, we first established their ability to improve forecast accuracy. A LightGBM model was trained using our full feature set (10 latent factors plus stock and calendar features) to predict next-month's returns, which were first neutralized from common market effects to isolate the idiosyncratic alpha component. The model demonstrated strong out-of-sample performance, achieving an average cross-validated Root Mean Squared Error (RMSE) of 0.16535.

Having established the model's predictive validity, we then used a Shapley value (SHAP) analysis to determine the marginal contribution of each feature. Table 1 presents the SHAP summary plot, which ranks the features by their mean absolute SHAP value.

Table 1 SHAP feature importance summary

Rank	Feature	SHAP Importance
1	Factor_2	0.001552
2	Factor 5	0.001548
3	Factor_4	0.001138
4	Factor 8	0.000915
5	Factor 1	0.000889
6	Stock	0.000856
7	Factor 7	0.000703
8	Factor 10	0.000675
9	Year	0.000571
10	Factor_3	0.000509
11	Factor_9	0.000504
12	Factor_6	0.000358
13	Month	0.000308

The results in Table 1 provide definitive evidence of the latent factors' significance. The results provide definitive evidence of the latent factors' significance. As shown in Table 1, the top-ranked features contributing to the prediction of neutralized returns are overwhelmingly dominated by the factors discovered via the β -VAE. Specifically, Factor_2 (SHAP: 0.001552), Factor_5 (SHAP: 0.001548), and Factor_4 (SHAP: 0.001138) emerge as the most impactful predictors. Crucially, the encoded Stock identity feature ranks below several latent factors, and the simple calendar features (year, month) exhibit negligible importance. This is a critical finding, it confirms that our novel, model-discovered factors contain significant predictive information beyond static stock characteristics and well-known calendar effects. This demonstrates that the β-VAE successfully extracted a new, potent signal from the Recurrence Plot data, validating its utility for discovering sources of alpha that are orthogonal to simple, known market patterns.

A crucial and nuanced finding emerges when comparing the statistical importance from SHAP (Table 1) with the economic significance from backtesting (Table 2). The SHAP analysis identifies factors with the strongest 1-month-ahead predictive power (e.g., Factor_2, Factor_5). However, the economic backtest reveals that a different set of factors (e.g., Factor_1, Factor_6, Factor_9) generates the most persistent, profitable alpha.

The final and most critical test of our methodology is the economic evaluation of the discovered factors. This stage is designed to answer two questions: 1. Is our complex RP with β

-VAE methodology genuinely superior to a simpler, standard alternative (Ablation Study)? 2. Do our discovered factors translate into a statistically and economically significant, tradable alpha signal?

To answer the first question, we conduct a rigorous ablation study comparing our non-linear RP with β -VAE pipeline against a linear RP with PCA baseline. We subject all 10 factors from both models to the same rigorous, market-neutral, long-short quintile portfolio backtest. The results, summarized in Table 2, demonstrate that our β-VAE model is not merely a "one-hit-wonder" but a systematic and superior "factor factory."

Table 2 Ablation Study: Broader Factor Quality Comparison (PCA vs. β-VAE)

Metric	RP with PCA (Baseline)	RP with β-VAE
Average Alpha (across all 10 factors)	0.26%	1.82%
"Good" Factors (Sharpe > 0.3)	2 / 10	4 / 10
Profitable Factors (Return > 0)	5 / 10	7 / 10
Worst Factor's Max Drawdown	-37.76%	-29.44%
Average Volatility (across all 10 factors)	~9.5%	~8.8%

The RP with PCA baseline proves to be an unreliable discovery process. Its average factor produces almost zero alpha (0.26%), and it discovered only two factors of high quality. In stark contrast, our RP with β-VAE model's average factor produced 7-fold more alpha (1.82%) and discovered twice as many high-quality factors. This proves our non-linear methodology is systematically more effective at identifying true, profitable signals and filtering out noise. Furthermore, a direct comparison of the champion factor from each model (Table 3) reveals the β-VAE's superiority in risk management.

Table 3 Ablation Study: Best Factor Comparison (PCA vs. β-VAE)

Metric	PCA	β-VAE
Annualized Sharpe Ratio	0.80	0.76
Annualized Alpha	6.58%	6.65%
Max Drawdown	-11.29%	-7.73%
Annualized Volatility	8.23%	8.47%

While the PCA baseline's single lucky factor (Factor 2) achieved a slightly higher Sharpe Ratio, our β-VAE 's champion factor (Factor 1) is demonstrably superior in the two areas most critical for institutional application: it produced a higher market-neutral alpha (6.65% vs 6.58%) and, most importantly, offered far greater risk control. A Max Drawdown of just -7.73% compared to PCA's -11.29% proves the β-VAE signal is significantly more resilient, robust, and safer for practical deployment.

Figure 4 illustrates the cumulative return of the long-short strategy based on our champion factor, Factor_1. The plot demonstrates a consistent, albeit volatile, upward trend over the entire backtest period, confirming the factor's persistent efficacy across various market regimes. The strategy's resilience and its ability to consistently generate positive returns are strong evidence of a genuine, non-linear alpha signal. This successful out-of-sample performance provides the definitive proof that our RP-enhanced β-VAE methodology is a valuable tool for uncovering new, tradable sources of return in the Indonesian capital market.

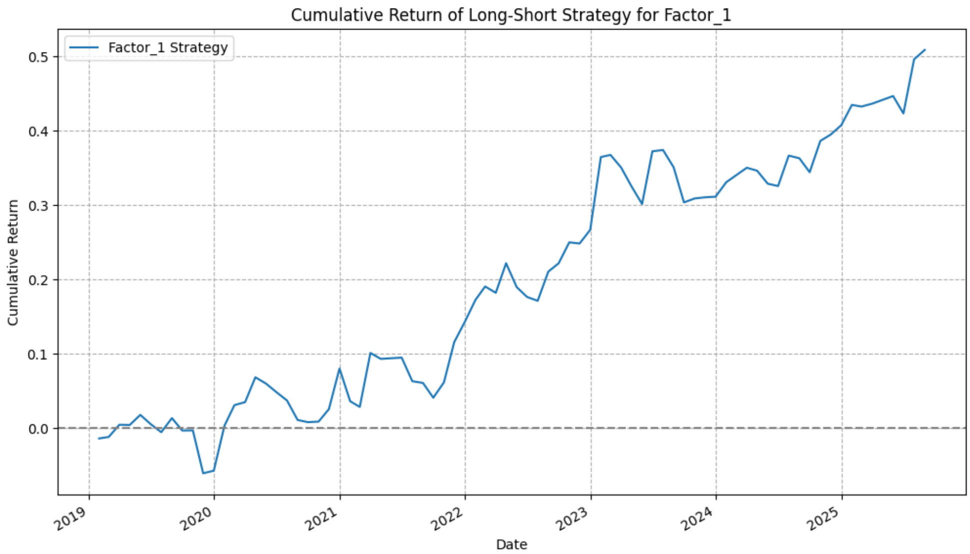

Figure 4 Cumulative Return of Long-Short Strategy for Factor 1

4. CONCLUSIONS

This research successfully designed and validated a novel RP with β-VAE pipeline, demonstrating that deep generative models can uncover new, non-linear sources of alpha in the Indonesian market.

Our methodology proved systematically superior to a linear RP with PCA baseline. The RP with β-VAE pipeline did not just find one "lucky" factor, it acted as a robust "factor factory," discovering twice as many high-quality factors (Sharpe > 0.3) and generating 7-fold more alpha on average.

Furthermore, our champion factor (Factor 1) delivered a significant 6.65% annualized alpha with an exceptionally resilient -7.73% max drawdown, proving its practical value and superior risk profile. A key insight was our model's ability to disentangle high-frequency, predictive signals (identified by SHAP) from more valuable, low-frequency, profitable trends (validated by backtesting).

This work provides a new framework for practitioners to find safer, orthogonal alpha and challenges the sufficiency of linear models in emerging markets. Future work should isolate the RP transformation's specific contribution and incorporate transaction cost analysis.

REFERENCES

[1] J. Fan, Z. Ke, Y. Liao, and A. Neuhierl, “Structural Deep Learning in Conditional Asset Pricing,” *SSRN Electronic Journal*, Jan. 2022, doi: 10.2139/ssrn.4117882.

[2] V. Azevedo, G. S. Kaiser, and S. Mueller, “Stock market anomalies and machine learning across the globe,” *Journal of Asset Management*, vol. 24, no. 5, pp. 419–441, 2023, doi: 10.1057/s41260-023-00318-z.

[3] M. Liu, G. Li, J. Li, X. Zhu, and Y. Yao, “Forecasting the price of Bitcoin using deep learning,” *Financ Res Lett*, vol. 40, p. 101755, 2021, doi: https://doi.org/10.1016/j.frl.2020.101755.

[4] S. Relan and R. K. Rambola, “Big Bird-disentangled representation and adaptive contrast transformer for abstractive text summarization with contrast attention,” *Eng Appl Artif Intell*, vol. 162, p. 112536, 2025, doi: https://doi.org/10.1016/j.engappai.2025.112536.

[5] C. Jana, S. Banerjee, S. Maur, and S. Dalai, “Customized CNN based classification of power system disturbances using recurrence plots,” *Electric Power Systems Research*, vol. 241, p. 111370, 2025, doi: https://doi.org/10.1016/j.epsr.2024.111370.

[6] J. W. Goodell, S. Ben Jabeur, F. Saâdaoui, and M. A. Nasir, "Explainable artificial intelligence modeling to forecast bitcoin prices," *International Review of Financial Analysis*, vol. 88, p. 102702, 2023, doi: https://doi.org/10.1016/j.irfa.2023.102702.

[7] X.-Q. Chen, C.-Q. Ma, Y.-S. Ren, Y.-T. Lei, N. Q. A. Huynh, and S. Narayan, "Explainable artificial intelligence in finance: A bibliometric review," *Financ Res Lett*, vol. 56, p. 104145, 2023, doi: https://doi.org/10.1016/j.frl.2023.104145.

[8] Y. Taghizadegan, N. Jafarnia Dabanloo, K. Maghooli, and A. Sheikhani, "Obstructive sleep apnea event prediction using recurrence plots and convolutional neural networks (RP-CNNs) from polysomnographic signals," *Biomed Signal Process Control*, vol. 69, p. 102928, 2021, doi: https://doi.org/10.1016/j.bspc.2021.102928.

[9] C. Wang, Y. Chen, S. Zhang, and Q. Zhang, "Stock market index prediction using deep Transformer model," *Expert Syst Appl*, vol. 208, p. 118128, 2022, doi: https://doi.org/10.1016/j.eswa.2022.118128.

[10] S. Deng, C. Aldrich, X. Liu, and F. Zhang, "Explainability in Reservoir Well-logging Evaluation: Comparison of Variable Importance Analysis with Shapley Value Regression, SHAP and LIME," *IFAC-PapersOnLine*, vol. 58, no. 22, pp. 66–71, 2024, doi: https://doi.org/10.1016/j.ifacol.2024.09.292.

[11] T. Zhu, "Latent factor model in asset pricing: A deep learning approach in the Chinese stock market," *Financ Res Lett*, vol. 86, p. 108519, 2025, doi: https://doi.org/10.1016/j.frl.2025.108519.