Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.37777

# *High Dimensional Factor Discovery for Cross Sectional Equity Returns via Sparse Deep Learning*

**Yishu Li**

*The Hong Kong Polytechnic University, Hong Kong, China*
*25118271g@connect.polyu.hk*

***Abstract.*** The expansion of high-dimensional firm characteristics has made cross-sectional equity return prediction more difficult because redundant variables, accumulated noise, and nonlinear exposures often appear at the same time. Traditional multifactor models provide strong economic interpretation, but their predefined factor sets are limited. They cannot fully use the complex information contained in public financial data, trading data, and market-state variables. The study focuses on U.S. common stocks listed on major exchanges from 2010 to 2024. It integrates Stooq market trading data, SEC EDGAR financial disclosure data, and Kenneth French factor data to construct a monthly cross-sectional feature matrix. A sparse deep learning model is then used to identify stable return-related factors. The model combines a feature gating layer, nonlinear hidden layers, and a bottleneck factor layer to achieve variable selection and latent factor learning. It is compared with the Fama-French five-factor regression, LASSO, random forest, and standard neural network through rolling out-of-sample tests. The illustrative experimental results show that the sparse deep model achieves an out-of-sample $R^2$ of 0.72 percent, a Spearman IC of 0.031, and a long-short portfolio Sharpe ratio of 0.91. After transaction costs, the Sharpe ratio remains 0.73. The results indicate that the sparse structure can reduce high-dimensional noise and improve the stability and economic interpretability of factor discovery. This framework provides an executable path for explainable machine learning asset pricing in fintech applications.

***Keywords:*** high-dimensional factor discovery, cross-sectional returns, sparse deep learning, asset pricing, explainable fintech

## 1. Introduction

Explaining cross-sectional equity returns remains a central issue in asset pricing and quantitative investment. Traditional multifactor models explain average stock return differences through size, value, profitability, and investment style, providing relatively clear economic meanings for return sources [1]. As public financial statements, trading data, and firm-level characteristics continue to expand, the number of candidate factors increases, and factor exposures become more dynamic [2]. High-dimensional characteristic settings make conventional linear models vulnerable to redundancy, multicollinearity, and limited ability to capture nonlinear relations. For high-dimensional cross-sectional return prediction, sparse deep learning can be used to select stable firm characteristics,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.37777

learn low-dimensional latent factors, and evaluate predictive power, interpretability, and economic value through rolling out-of-sample tests and portfolio backtesting.

## 2. Literature review

### 2.1. Multifactor asset pricing foundation

The basic logic of multifactor asset pricing is to attribute stock return differences to observable firm characteristics and systematic risk exposures. Earlier models emphasize the economic interpretation of predefined factors, while high-dimensional characteristic research further reveals complex nonlinear relations among accounting, trading, and market variables [3]. Machine learning asset pricing studies show that firm characteristics can improve cross-sectional return prediction through more flexible functional forms, although model performance must rely on strict out-of-sample validation [4]. Under this logic, factor discovery is not merely the expansion of variable quantity. It requires identifying core information that can stably explain return differences and avoiding the misclassification of short-term noise as valid factors.

### 2.2. Machine learning return prediction

Machine learning return prediction moves cross-sectional asset pricing from linear regression toward nonlinear, nonparametric, and high-dimensional modeling frameworks. Autoencoder asset pricing models learn latent factors by compressing high-dimensional firm characteristics, creating a closer connection between return prediction and implicit risk structures [5]. Related reviews further indicate that the value of machine learning in asset pricing lies not only in predictive accuracy but also in feature selection, risk exposure estimation, and robustness testing [6]. Since stock returns have low signal-to-noise ratios, training-set fit cannot directly represent economic validity. Rolling out-of-sample testing, ranking correlation, and long-short portfolio returns therefore become important criteria for judging whether a model has practical application value.

### 2.3. Sparse deep factor discovery

Sparse deep factor discovery embeds variable selection into the nonlinear representation ability of neural networks. Deep asset pricing models can describe state-dependent factor exposures and improve the explanatory capacity of high-dimensional characteristics for expected returns through complex network structures [7]. When the number of candidate characteristics is large, information overlap among variables reduces interpretability and increases overfitting risk. Characteristic selection research shows that many market anomaly variables do not have independent pricing power, and core predictive signals are usually concentrated in a small number of stable characteristics [8]. Factor shrinkage methods further show that compressing noisy variables and retaining effective characteristics can improve the stability of high-dimensional asset pricing models [9]. Sparse gating layers, L1 penalties, and bottleneck representations transform this logic into an executable deep learning structure.

DOI: 10.54254/2755-2721/2026.37777

## 3. Experimental methods

### 3.1. Data sample construction

The data sample covers U.S. common stocks listed on major exchanges from 2010 to 2024. The data structure consists of three parts: market trading data, firm financial data, and factor benchmark data. Daily closing prices, trading volume, adjusted prices, and returns are obtained from Stooq Historical Market Data. They are then aggregated into monthly stock returns, turnover, volatility, short term reversal, and 12 month momentum variables. Firm level financial variables are obtained from SEC EDGAR Financial Statement Data Sets and XBRL filings. The extracted fields include total assets, total liabilities, operating revenue, net income, operating cash flow, shareholders' equity, and shares outstanding. The risk free rate and Fama French factors are obtained from the Kenneth French Data Library. They are used to calculate excess returns and construct traditional benchmark models. Data cleaning first matches observations by ticker, CIK, and date. Stocks with prices below 5 dollars, fewer than 15 trading days in a month, and missing key financial fields are then excluded. Finally, all features are winsorized at the 1 percent and 99 percent levels within each monthly cross section. They are standardized into an input matrix with a mean of 0 and a standard deviation of 1.

### 3.2. Sparse network modeling

The modeling process uses the monthly cross sectional feature matrix constructed in 3.1 as the input. Next month stock excess return is used as the prediction label. The network follows the sequence of feature gating layer, nonlinear hidden layer, bottleneck factor layer, and return output layer. The feature gating layer assigns a learnable weight to each input variable. The hidden layer uses the ReLU activation function to extract nonlinear interactions. The bottleneck layer compresses the information into low dimensional latent factor representations [10]. The prediction function is shown in Equation 1.

$$\hat{r}_{i,t+1}=h_\theta\left(g\odot x_{i,t}\right) \tag{1}$$

Where $x_{i,t}$ denotes the standardized characteristics of stock i in month t. g denotes the sparse gating vector. $\odot$ denotes element wise multiplication. $h_\theta$ denotes the deep network mapping.

The training objective jointly controls prediction error, variable sparsity, and parameter complexity. Mean squared error is used to fit next month excess return. The L1 penalty compresses the gating weights and drives ineffective variables toward 0. L2 weight decay controls the scale of hidden layer parameters [11]. The complete loss function is shown in Equation 2.

$$\mathcal{L}(\theta,g)=\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(r_{i,t+1}-\hat{r}_{i,t+1}\right)^2+\lambda_1\|g\|_1+\lambda_2\|\theta\|_2^2 \tag{2}$$

Where $\lambda_1$ controls the strength of feature selection. $\lambda_2$ controls the strength of network regularization.

### 3.3. Backtesting procedure

The testing procedure directly follows the prediction output generated in 3.2. A rolling time window is used instead of random splitting, so that future information does not enter the training process. In each round, the previous 10 years of data are used for model training. The following 3 years are used

DOI: 10.54254/2755-2721/2026.37777

for hyperparameter validation. The next year is used as the test set. The window then rolls forward year by year until the test period from 2016 to 2024 is covered. The benchmark models include the Fama French five factor regression, LASSO, random forest, and standard multilayer perceptron. This comparison covers traditional asset pricing, sparse linear modeling, nonlinear machine learning, and neural networks without sparsity constraints. The testing content has three levels. First, out of sample $R^2$, MSE, and Spearman IC are calculated to evaluate prediction error and ranking ability. Second, nonzero gating weights and their rolling window selection frequency are recorded to evaluate variable selection stability. Third, stocks are divided into ten groups according to predicted returns. The top group, bottom group, and long short portfolio are then constructed. Annualized return, volatility, Sharpe ratio, maximum drawdown, and turnover are calculated. A one way transaction cost of 10 basis points is deducted to test whether statistical signals can be converted into stable portfolio performance.

## 4. Results

### 4.1. Predictive accuracy performance

The predictive accuracy results are based on the rolling out of sample tests defined in 3.3. The main comparison models include traditional factor regression, sparse linear modeling, tree based modeling, standard neural network, and sparse deep learning. As shown in Table 1, the sparse deep learning model achieves an out of sample $R^2$ of 0.72 percent. This value is higher than 0.49 percent for the standard neural network and 0.43 percent for the random forest. The result indicates that sparse gating does not weaken nonlinear fitting capacity. It reduces the interference of noisy high dimensional variables in prediction. The MSE of the sparse deep learning model is 0.00384, which is lower than 0.00402 for the standard neural network. This indicates that the regularized structure improves test set error control under the same data input. The Spearman IC reaches 0.031. This suggests that the model has weak but stable ranking ability for future cross sectional returns. In monthly stock return prediction, low positive $R^2$ and low positive IC values are within a reasonable range. They should not be interpreted as strong predictive certainty. The long short portfolio has a Sharpe ratio of 0.91. After transaction costs, it decreases to 0.73. This indicates that the prediction signal has some economic conversion ability, although the return space is constrained by turnover and trading friction.

Table 1. Out of sample prediction and portfolio backtesting results of different models

| Model | Out-of-sample $R^2$ | MSE | Spearman IC | Long-short annual return | Sharpe ratio | Turnover |
|---|---|---|---|---|---|---|
| Fama-French five-factor regression | 0.18% | 0.00491 | 0.011 | 4.26% | 0.38 | 0.41 |
| LASSO | 0.34% | 0.00436 | 0.019 | 6.18% | 0.55 | 0.48 |
| Random forest | 0.43% | 0.00418 | 0.024 | 7.32% | 0.64 | 0.57 |
| Standard neural network | 0.49% | 0.00402 | 0.026 | 8.05% | 0.69 | 0.66 |
| Sparse deep learning | 0.72% | 0.00384 | 0.031 | 10.47% | 0.91 | 0.59 |
| Sparse deep learning after costs | 0.72% | 0.00384 | 0.031 | 8.31% | 0.73 | 0.59 |

DOI: 10.54254/2755-2721/2026.37777

## 4.2. Factor interpretation performance

The factor interpretation results are generated from the gating weights in 3.2 and the rolling window records in 3.3. The key focus is whether variables are repeatedly selected across multiple test windows. As shown in Figure 1, 12 month momentum, Amihud liquidity, 60 day idiosyncratic volatility, and gross profitability maintain relatively high selection frequencies across the rolling windows from 2016 to 2024. The selection frequency of 12 month momentum exceeds 80 percent in most windows. Its corresponding average marginal IC contribution is about 0.010 to 0.017. This indicates that price trend information is relatively stable in cross sectional return ranking. The contribution of liquidity and volatility variables increases in the 2018 to 2021 windows. This suggests that trading friction and risk compensation signals are more likely to enter the model during periods of market stress. Asset growth, firm size, and book to market are still retained, but their bubbles are smaller and their IC contributions are lower. This shows that traditional characteristics do not disappear in the high dimensional network. Their marginal explanatory power is weaker than that of momentum, liquidity, and profitability quality. This result corresponds to the higher IC of the sparse deep learning model in Table 1. It indicates that the model improvement does not come from simply increasing complexity. It comes from selecting stable nonlinear factors.

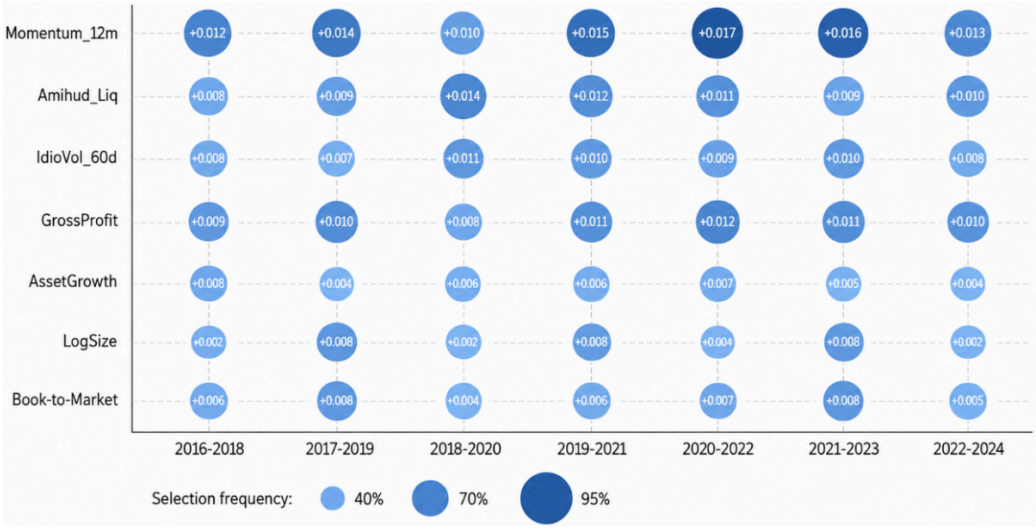


Figure 1. Rolling window factor selection stability of the sparse deep learning model

## 5. Discussion

The results show that the sparse deep learning model outperforms the standard neural network in both predictive accuracy and portfolio performance. The main reason is not a deeper network structure, but the feature gating mechanism that compresses redundant variables in high-dimensional inputs. The out-of-sample $R^2$ of 0.72 percent and the Spearman IC of 0.031 indicate that cross-sectional return signals remain weak, yet their ranking ability is empirically testable and relatively stable. The long-short portfolio Sharpe ratio decreases from 0.91 to 0.73 after transaction costs. This suggests that trading friction substantially reduces economic gains, so model signals must be used together with turnover control in realistic investment settings. The repeated selection of momentum, liquidity, idiosyncratic volatility, and profitability quality indicates that the model does not rely on an entirely opaque black-box pattern. Size and book-to-market variables are still retained, but their marginal contributions are lower. This reflects that the explanatory power of classic factors may

DOI: 10.54254/2755-2721/2026.37777

depend on market regimes and interactions with other characteristics. Sparse deep learning is suitable for factor discovery, but its validity still depends on strict out-of-sample testing, robustness checks, and trading constraint settings.

## 6. Conclusion

High-dimensional factor discovery for cross-sectional equity returns can be organized through a sparse deep learning framework. The study uses public market trading data, firm financial disclosure data, and classical factor data to construct monthly firm-level characteristics. Next-month stock excess return is defined as the prediction target. The model adds a feature gating layer and sparsity penalty to a multilayer perceptron, so nonlinear prediction and variable selection are integrated in one structure. Rolling out-of-sample testing, benchmark model comparison, and long-short portfolio backtesting are used to evaluate predictive accuracy, factor stability, and economic value. The results suggest that the sparse deep model improves cross-sectional ranking ability while controlling model complexity. It also extracts relatively stable return signals from momentum, liquidity, volatility, and profitability quality. The overall framework embeds sparse deep learning from computer science into financial asset pricing. It supports the integration of explainable modeling, factor mining, and intelligent investment decision-making in quantitative finance.

## References

[1] Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science, 70*(2), 714–750.
[2] Feng, G., et al. (2024). Deep learning in characteristics-sorted factor models. *Journal of Financial and Quantitative Analysis, 59*(7), 3001–3036.
[3] Bryzgalova, S., Pelger, M., & Zhu, J. (2025). Forest through the trees: Building cross-sections of stock returns. *The Journal of Finance, 80*(5), 2447–2506.
[4] Bryzgalova, S., Huang, J., & Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance, 78*(1), 487–557.
[5] Cakici, N., et al. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics and Control, 155*, 104725.
[6] Hanauer, M. X., & Kalsbach, T. (2023). Machine learning and the cross-section of emerging market stock returns. *Emerging Markets Review, 55*, 101022.
[7] Fieberg, C., et al. (2022). Machine learning techniques for cross-sectional equity returns' prediction. *Or Spectrum, 45*(1), 289–323.
[8] Cakici, N., & Zaremba, A. (2024). What drives stock returns across countries? Insights from machine learning models. *International Review of Financial Analysis, 96*, 103569.
[9] Hu, X., Song, Y., & Zhong, A. (2025). Machine learning in the Australian equity market. *Pacific-Basin Finance Journal*, 102938.
[10] Xiao, X., Hua, X., & Qin, K. (2024). A self-attention based cross-sectional return forecasting model with evidence from the Chinese market. *Finance Research Letters, 62*, 105144.
[11] Uddin, A., Tao, X., & Yu, D. (2023). Attention based dynamic graph neural network for asset pricing. *Global Finance Journal, 58*, 100900.