**Online Appendices**

This online appendix accompanies the manuscript titled: "**A Systematic Literature Review of Asset Pricing: Insights from AI and Big Data**".

**Contents**

## Appendix S1: Methodological Review and Data Scope of Included Papers

**Table S1**: Methodological Review of AI in Asset Pricing

| Reference | AI Model | Theoretical and Empirical Framework | Functionality of AI Technique | Contribution |
|---|---|---|---|---|
| Akyilidirim et al. (2021) | Multiple Machine Learning models | | | Prediction accuracy of AI models |
| Alaminos et al. (2023) | Artificial Neural networks (Multilayer Perceptron (DRCNN), Deep Neural Decision Trees (DNDT), Quantum Neural Network) | | Nonlinear function approximations | A model that captures the pricing behaviour of football clubs efficiently |
| Alex Avelar et al. (2024) | Multiple Machine Learning models | | | AI models efficiently exploit market efficiency |
| Alves et al. (2022) | Least Absolute Shrinkage and Selection Operator (LASSO) and Adaptive LASSO | Factor model | Regularisation and Dimension reduction | The model improves forecasting accuracy |
| Anis and Kwon (2021) | Multiple Machine Learning models | Factor model | Dimension reduction and Nonlinear function approximation | A model that examines the process of constructing factor models effectively |
| Auh and Cho (2023) | Structural Vector Regression | Factor model | Noise filtering and Nonlinear function approximation | Optimised factor analysis |
| Azevedo et al. (2023) | Multiple AI models | Factor model | Feature recognition and Nonlinear function approximation | Feedforward neural network has the highest return predictability |
| Babikir and Mwambi (2017) | Artificial neural network (FAANN) | Factor model | Nonlinear function approximation | The model offers forecast accuracy |
| Bang (2024) | Double-selection LASSO | Factor model | Dimension reduction | Suggests that significant factors can vary across markets |
| Begušić et al. (2020) | Multiple Machine Learning models | Factor model | Dimension reduction and Feature extraction | The models yield accurate clustering results |
| Cakici et al. (2023) | Machine learning models (LASSO, Elastic net, support vector machine (SVM), Gradient boosted regression trees (GBRT), Random forests (RF), Feed-forward eural networks with one, two, or three hidden layers (NN1, NN2, NN3)) | Ordinary least squares regressions (OLS), Partial least squares (PLS) | | The forecast combination model is superior to individual models.<br>The effectiveness of machine learning models is not equal |
| Cao (2011) | Multiple Artificial Neural Network | CAPM and Factor model | Nonlinear function approximation | The ANN-based models improve forecasting accuracy |
| Cao et al. (2005) | Multiple Artificial Neural Network | CAPM and Factor model | Nonlinear function approximation | The results show that neural networks outperform the linear models compared |
| Cerqueti et al. (2024) | Autoencoder Neural Network | Factor model | Dimension reduction and Nonlinear function approximation | The model captures complex relationships and nonlinearities while estimating latent factors which explain commodity prices |
| L. Y. Chen et al. (2024) | Deep Neural Networks (Feedforward network, Long- and Short-term memory (LSTM), GAN) | Factor model | Nonlinear function approximation, temporal and generative modelling | A nonlinear asset pricing model that extracts economic conditions from complex data |
| Y. C. Chen et al. (2022) | Back-propagation neural network (BPNN) | | Error minimisation, Feature extraction and Nonlinear function approximation | A model that adapts to complex conditions and has good forecasting ability |
| Costa et al. (2021) | Multiple AI models | | Ensemble learning, Regularisation, Supervised learning | AI models outperform traditional asset pricing models |
| Cuomo et al. (2022) | Machine Learning models (LASSO, Neural Network Principal Component Analysis and Variational autoencoders) | Arbitrage Pricing Theory | Data clustering, Feature extraction and Nonlinear function approximation | The model optimises portfolio strategy |
| De Nard et al. (2021) | Multiple Machine Learning models | | Dimension reduction and Ensemble learning | Variable subsample aggregation performs well under different levels of sparsity and signal-to-noise ratios |
| Diallo et al. (2022) | Support Vector Regression (SVR) | Factor model | Nonlinear function approximation | Accurate predictions of portfolio returns |

| | | | | |
|---|---|---|---|---|
| Dixon and Polson (2020) | Deep Neural Networks | Factor model | Nonlinear functions and Interaction modelling | A model that generalises linear fundamental factor models |
| Erfanian et al. (2022) | Ensemble Learning, Support Vector Regression, Multilayer Perception | | Error minimisation, Feature extraction and Nonlinear function approximation | Researchers doubt that machine learning can beat the traditional methods in Bitcoin price prediction |
| Fan et al. (2017) | Deep Learning Model | Factor model | Dimension reduction, Feature extraction and Regularisation | A model that reduces dimensionality and can extract predictive indices |
| Fang and Taylor (2021) | Multiple AI models | Factor model | | Increased predictive performance of linear regressions fed into AI models |
| Fang et al. (2023) | Support Vector Regression (SVR) | Factor model/ Discrete Wavelet Transform | Feature extraction and Nonlinear function approximation | Incorporating machine learning into factor models improves their pricing power significantly |
| Feng et al. (2024) | Deep Learning Model | Factor model | Dimension reduction and Nonlinear function approximation | A model for constructing latent factors for asset pricing |
| Gu et al. (2020) | Multiple AI models | | Regularisation, Dimension reduction, Nonlinear functions | Provide benchmarks for assessing the predictive accuracy of machine learning techniques in measuring risk premiums |
| Gu et al. (2021) | Autoencoder (Neural Networks) | Factor model | Dimension reduction | A nonlinear conditional asset pricing model that can identify latent factors |
| Han (2021) | Machine Learning Models (Sparse Multinomial Logistic Regression (SMLR) – Fast Parallel SMLR, Spectral Spatial SMLR) | Factor model | Learning optimisation and Feature extraction | The model can explain factors affecting asset prices |
| Huang et al. (2022) | Deep Learning Model (LSTM and Matrix-based feature fusion) | | Feature interactions and Nonlinear function approximation | The model allows the simultaneous estimation of the intensities of feature interactions and firm interferences |
| Healy et al. (2024) | Automated Machine Learning model | Factor model | Nonlinear function approximation | Machine learning methods provide more accurate models for predicting stock returns based on risk factors than standard regression-based methods of estimation |
| Huang (2022) | XGBoost | Factor model | Feature extraction and Pattern recognition | Prediction accuracy of AI models |
| Jan and Ayub (2019) | Artificial Neural Networks | Factor model | Nonlinear function approximation | Accurately predicts returns in stock markets |
| Khoa and Huynh (2023) | Support Vector Regression (SVR) | CAPM and Factor model | Nonlinear function approximation | The 3-factor model integrated into SVR provides superior explanation for stock returns |
| Khoa and Huynh (2022) | LSTM Recurrent Neural Network (LSTM-RNN) | Factor model | Nonlinear function approximation | LSTM-RNN improves the accuracy of 5FF model and effectively exploits latent factors |
| Kılıc et al. (2023) | Multiple AI models | | | AI models can predict the direction of intraday price movements |
| Kim et al. (2023) | Conditional Autoencoder | Factor model | Feature extraction and Nonlinear function approximation | The models show excellent explanatory performance |
| Lalwani and Meshram (2022) | Machine Learning models (Artificial Neural Network, Elastic Net, Random forests, XGBoost) | | Nonlinear function approximation | Machine learning models offer accurate forecasts |
| Li and Mei (2020) | Deep Learning Model | Standard linear models (ordinary least squares, stepwise regression, principal components regression, elastic net) | Regularisation and Nonlinear function approximation | The application of deep learning methods to asset returns can yield more intuitive results than those derived using standard methods |
| Liu et al. (2023) | Autoencoder (Feedforward Neural Network) | Factor model | Regularisation | A model that captures nonlinear relationships in factors |
| Lo and Singh (2023) | Deep Neural Network | | Feature extraction and Nonlinear function approximation | The model outperforms other machine learning models |
| Loo (2019) | Artificial Neural networks | CAPM and Factor model | Nonlinear function approximation | ANN produced better forecasting results than the regression models |
| Ma, Liao et. (2023) | Deep Learning Model | Factor model | Dimension reduction and Nonlinear function approximation | The model has accurate forecasts, higher Sharpe ratios and better certainty equivalent returns |

| | | | | |
|---|---|---|---|---|
| Maasoumi et al. (2022) | Automatic debiased machine learning (ADML) | Factor model | Regularisation | The model enhances accuracy and supports stable predictions |
| Maeshashi and Shintani (2020) | Multiple AI models | Factor model | Ensemble learning, Nonlinear function approximation, and Interaction effect | Machine learning models are accurate for predictions |
| Martínez et al. (2020) | Multiple Machine Learning models | | | AI models efficiently exploit market efficiency |
| Massa Roldán et al. (2022) | LSTM | Wavelet | Noise reduction, Signal decomposition and Nonlinear function approximation | An advanced model useful for accurate high-frequency predictions |
| Pan et al. (2023) | LSTM Neural Network | Factor model | Nonlinear function approximation and temporal deep learning | A model that captures nonlinear long-short-term memory structure |
| Park et al. (2024) | Multiple Machine Learning models | Factor model | Classification and Nonlinear function approximation | |
| Refenes et al. (1993) | Neural Networks | Arbitrage Pricing Theory | Nonlinear function approximation | Neural network models accurately predict returns |
| Rui (2023) | Machine Learning Model | Factor model | Transfer learning | The proposed model is useful for analysing factors and predictions |
| Tang et al. (2023) | Multiple Machine Learning Models | Factor model | Nonlinear functions and Feature selection | DeepForest-CQP multi-factor model constructs a stock selection that achieves higher returns |
| Uddin and Yu (2020) | Autoencoder Neural Network | Factor model | Feature extraction/ Pattern recognition | A model that effectively extracts latent factors for explaining and predicting risk premia |
| Uddin et al. (2023) | Machine learning models (Graph learning Spectral and temporal dynamics learning) | | Dimensional reduction, Embedded learning, Graph learning and Nonlinear functions | The proposed model effectively detects significant influences in a firm network and models how inform spreads in the network |
| Wang (2024) | Multiple AI models | Factor model | Dimensional reduction, Nonlinear functions and Interaction modelling | Neural network models effectively predict returns |
| Wang (2021) | Multiple Machine Learning models | Factor model | Regularisation, Classification and Nonlinear function approximation | Machine learning models could explain most of the excess return of cryptocurrencies |
| X. L. Yang et al. (2024) | Autoencoder | Factor model | Generative and deep learning | A model that estimates the conditional distribution of returns for asset pricing |
| J. H. Yang et al. (2024) | Hierarchical Deep Learning Model | | Nonlinear function approximation and Feature extraction | Reliance on CAPM, herding effect, holiday effect may not capture all market dynamics |
| Yao et al. (2022) | LSTM Neural Network | Factor model | Nonlinear functions and Feature extraction | The 6-factor LSTM model successfully forecasts asset price using a limited number of high-quality pricing factors |
| Yao et al. (2024) | LSTM Neural Network | Factor model | Nonlinear functions and Feature extraction | The proposed model has an enhanced model interpretability that consistently delivers better performance |
| Yuyan et al. (2023) | Multiple Machine Learning models | Volatility models (heterogeneous autoregressive quarticity (HARQ) and full HARQ) | Supervised learning | Hybrid volatility models improve predictions |
| Zhan et al. (2022) | Multiple AI models | | | AI models can predict arbitrate opportunities |
| H. W. Zhang et al. (2023) | LASSO, Elastic Net, Gradient boosting, Random forests | Volatility models (heterogeneous autoregressive (HAR) | Regularisation, Dimension reduction, and Nonlinear function approximation | Machine learning models with higher-order moments effectively capture the trends in oil futures and the combination forecasting models have satisfactory results |
| R. X. Zhang et al. (2023) | Neural Network | | Image-based learning | A model that extracts information from images for predicting asset price |
| Zhu et al. (2020) | Machine Learning Model (Groupwise Interpretable Basis Selection (GIBS)) | Factor model | Regularisation and Dimension reduction | A model that effectively explains asset returns |
| Zhu et al. (2021) | Machine Learning Model (Groupwise Interpretable Basis Selection (GIBS)) | Factor model | Regularisation and Dimension reduction | The proposed model is more consistent with realised asset returns than factor models |

**Table S2**: Country and Data Scope

| Country | Data Scope | No of Articles: AI and Asset Pricing | Big data and Asset pricing | Big data, AI and Asset pricing |
|---|---|---|---|---|
| **China** | A-share stocks (Shanghai and Shenzhen stock markets) | 5 | 4 | 3 |
| | Public corporations data (Shanghai stock exchange) | 1 | 0 | 0 |
| | CSI 300 Index (Shanghai and Shenzhen stock markets) | 2 | 1 | 0 |
| | CSI 500 Index (Shanghai and Shenzhen stock markets) | 1 | 0 | 0 |
| | REITs (Hong Kong stock exchange) | 1 | 0 | 0 |
| | Chinese stock market indices and data | 4 | 2 | 0 |
| | Oil futures | 0 | 1 | 1 |
| **Europe** | Football club data | 1 | 1 | 0 |
| **India** | Stock data (Bombay and National stock exchange) | 1 | 0 | 0 |
| **Italy** | Stock data, forex and cryptocurrency data | 1 | 0 | 0 |
| **Japan** | Macroeconomic data | 1 | 0 | 0 |
| **Korea** | Stock data (KOSPI and KOSPI) | 3 | 0 | 0 |
| **Mexico** | Stock prices (Bolsa Mexicana de Valores) | 0 | 0 | 1 |
| **Pakistan** | Non-financial listed companies | 1 | 0 | 0 |
| | Stock data (Pakistan stock exchange) | 1 | 0 | 0 |
| **South Africa** | Deposit rate, interest rate and share prices | 1 | 0 | 0 |
| **Spain** | IBEX 35 | 0 | 1 | 0 |
| **Taiwan** | Trading companies | 1 | 0 | 0 |
| | Semi-conductor companies | 1 | 0 | 0 |
| | Optical lens manufacturing company | 0 | 1 | 1 |
| **Turkey** | Stock data | 1 | 2 | 1 |
| **UK** | Data from financial statements | 1 | 0 | 0 |
| **US** | S&P500 Index | 7 | 11 | 1 |
| | EFTs (NASDAQ, NYSE, DJI, CRSP) | 4 | 0 | 0 |
| | Macroeconomic data (CRSP and FRED-MD) | 4 | 6 | 3 |
| | Russell 3000 | 1 | 0 | 0 |
| | Russell 1000 | 1 | 0 | 0 |
| | Biotechnology companies | 1 | 1 | 1 |
| | DJI | 2 | 0 | 0 |
| | US stock data (multiple data scope – e.g., NYSE, NASDAQ, AMEX, S&P100, S&P500) | 12 | 10 | 6 |
| | S&P100 Index | 0 | 3 | 0 |
| | IBM stocks | 0 | 1 | 0 |
| | FTSE100 | 0 | 1 | 0 |
| **Vietnam** | Listed companies (Ho Chi Minh City stock market) | 1 | 0 | 0 |
| | Factor data (Hanoi stock market) | 1 | 0 | 0 |
| **Others** | Factor Data | 3 | 0 | 0 |
| | Global Data | 4 | 2 | 2 |
| | Cryptocurrency | 2 | 1 | 0 |
| | Commodities | 1 | 2 | 1 |
| | Review papers | 8 | 2 | 3 |
| | Theory and Synthetic data | 2 | 2 | 1 |
| | | **82** | **55** | **25** |

*Note*: This table presents the data scope in the included papers. Papers that use data across countries are captured under “global data”. Papers that developed theoretical models and inferences, used synthetic data or were silent about the data are captured under “theory and synthetic data”.

## Appendix S2: Bibliometric Analysis

**Table S3:** Bibliometric Presentation of the Most Productive Countries and Publication Sources

| Corresponding Author(s) Country | No of Articles | Total Citations/ Country | Publication Sources | No of Articles |
|---|---|---|---|---|
| | | **AI and Asset Pricing** | | |
| USA | 25 | 2096 | Journal of Econometrics | 4 |
| China | 21 | 216 | The Review of Financial Studies | 4 |
| Korea | 4 | 10 | Journal of Forecasting | 3 |
| France | 3 | 22 | Economics Letters | 2 |
| United Kingdom | 3 | 19 | Expert Systems with Applications | 2 |
| Australia | 2 | 31 | Finance Research Letters | 2 |
| Germany | 2 | 15 | IEEE Access | 2 |
| Italy | 2 | 13 | Journal of Computational Science | 2 |
| Pakistan | 2 | 11 | Journal of Economic Dynamics and Control | 2 |
| Spain | 2 | 5 | Journal of Intelligent and Fuzzy systems | 2 |
| Switzerland | 2 | 25 | Journal of Risk and Financial Management | 2 |
| Vietnam | 2 | 6 | Management Science | 2 |
| | | | Quantitative Finance | 2 |
| | | | Quarterly Journal of Finance | 2 |
| | | **Big Data and Asset Pricing** | | |
| China | 20 | 227 | Journal of Econometrics | 13 |
| USA | 17 | 3889 | Journal of Business and Economic Statistics | 4 |
| Spain | 4 | 55 | Management Science | 3 |
| United Kingdom | 3 | 9 | Econometrica | 2 |
| Brazil | 2 | 35 | International Journal of Forecasting | 2 |
| France | 2 | 12 | International Review of Financial Analysis | 2 |
| Germany | 2 | 26 | Journal of Financial Econometrics | 2 |
| Canada | 1 | 57 | Journal of Risk and Financial Management | 2 |
| Ireland | 1 | 8 | | |
| Switzerland | 1 | 24 | | |
| Denmark | 1 | 0 | | |
| | | **Big Data, AI and Asset pricing** | | |
| USA | 9 | 1652 | Journal of Econometrics | 2 |
| China | 8 | 34 | The Review of Financial Studies | 2 |
| Brazil | 2 | 35 | Ain Shams Engineering Journal | 1 |
| Canada | 1 | 57 | Applied Stochastic Models in Business and Industry | 1 |
| France | 1 | 8 | Connection Science | 1 |
| Ireland | 1 | 8 | Energy Economics | 1 |
| Italy | 1 | 14 | European Financial Management | 1 |
| Switzerland | 1 | 24 | European Journal of Finance | 1 |
| | | | Financial Markets and Portfolio Management | 1 |
| | | | International Journal of Financial Studies | 1 |

*Note*: The table presents the most productive countries and the number of citations for papers in these countries over the years. The most relevant sources for "AI and Asset Pricing" and "Big Data and Asset Pricing" with at least 2 publications are presented. For Big Data, AI and Asset pricing, the most relevant source with at least 2 publications is Journal of Econometrics only. Thus, the first 10 most relevant sources are presented.

**Table S4**: Bibliographic Coupling and Co-citation Analysis of AI and Asset Pricing

**Panel A: Bibliographic Coupling**

| Cluster (Theme) | Reference |
|---|---|
| 1 (AI models, high-frequency data approaches and asset pricing dynamics) | • Cao et al. (2005)<br>• Cao et al. (2011)<br>• Caporale et al. (2016)<br>• Erfanian et al. (2022)<br>• Gu et al. (2020)<br>• Ho et al. (2011)<br>• Huang and Hueng (2008)<br>• Huang et al. (2022)<br>• Panchenko et al. (2010)<br>• Refenes et al. (1994)<br>• Warin and Stojkov (2021)<br>• Weller (2018)<br>• Weigand (2019)<br>• Yuan and Lee (2015) |
| 2 (AI models for enhanced asset pricing) | • Bagnara (2024<br>• Cakici et al. (2023)<br>• L. Y. Chen et al. (2024)<br>• Feng et al. (2024)<br>• Giglio et al. (2022)<br>• Karolyi and Van Nieuwerburgh (2020) |
| 3 (AI models and factor modelling) | • Bai and Ng (2019)<br>• Fan et al. (2017)<br>• Giglio et al. (2021)<br>• Gu et al. (2021)<br>• Maehashi and Shintani (2020) |

**Panel B: Co-Citation Analysis**

| Cluster (Theme) | Cited Refence | Total Citations |
|---|---|---|
| 1 (Evolution of factor models) | • Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of finance, 66(4), 1047-1108. | 10 |
| | • Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. The Journal of Finance, 75(3), 1327-1370. | 11 |
| | • Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. The Review of Financial Studies, 33(5), 2326-2377. | 18 |
| | • Green, J., Hand, J. R., & Zhang, X. F. (2017). The characteristics that provide independent information about average US monthly stock returns. *The Review of Financial Studies*, *30*(12), 4389-4436. | 10 |
| | • Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273. | 36 |
| | • Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. Journal of Econometrics, 222(1), 429-450. | 18 |
| | • Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. The Review of Financial Studies, 29(1), 5-68. | 18 |
| | • Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics, 134(3), 501-524. | 14 |
| | • Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. Journal of Financial Economics, 135(2), 271-292. | 15 |
| | • Lettau, M., & Pelger, M. (2020). Factors that fit the time series and cross-section of stock returns. The Review of Financial Studies, 33(5), 2274-2325. | 11 |
| | • Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review 4 (1): 1–44*. | 10 |
| 2 (Asset pricing theories and factor analysis) | • Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191-221. | 13 |
| | • Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. the Journal of Finance, 47(2), 427-465. | 21 |
| | • Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. Journal of financial economics, 33(1), 3-56. | 37 |
| | • Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1), 1-22. | 31 |
| | • Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The journal of finance*, *20*(4), 587-615. | 14 |

| | | | | |
|---|---|---|---|---|
| | • Uddin and Yu (2020) | | • Ross, S. A. (2013). The arbitrage theory of capital asset pricing. In Handbook of the fundamentals of financial decision making: Part I (pp. 11-30). | 14 |
| 4<br>(Deep learning for enhanced financial modelling and risk prediction) | • Anis and Kwon (2021)<br>• Heaton et al. (2017)<br>• Luo et al. (2024) | | • Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442. | 25 |
| 5<br>(Machine learning and factor modelling) | • Jan and Ayub (2019) | 3<br>(Empirical tests and validation of classical asset pricing models) | • Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57-82. | 17 |
| | | | • Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3), 607-636. | 15 |
| | | | • Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3-12. | 13 |
| | | | • Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. The Review of Financial Studies, 28(3), 650-705. | 15 |
| | | | • Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of financial economics*, *108*(1), 1-28. | 10 |

*Note:* a.) Bibliographic coupling connects papers that share (same) references by grouping them into clusters. Shows papers that cite the same references (papers are used as unit of analysis). Thus, papers are assumed to be communicating the same idea based on the shared references. It is a backward citation analysis that shows references cited by 2 or more papers in the data corpus (in this case, a minimum of 10 shared references). There are 30 papers that share a minimum of 10 shared cited references out of the 82 papers.
b.) The co-citation analysis links papers that are frequently cited together by other papers. A cluster shows references that are usually cited today, reflecting their influence in the field (to identify key foundational works). Using a minimum of 10 citations of a cited reference, only 23 cited references meet this threshold out of 2857 cited references.

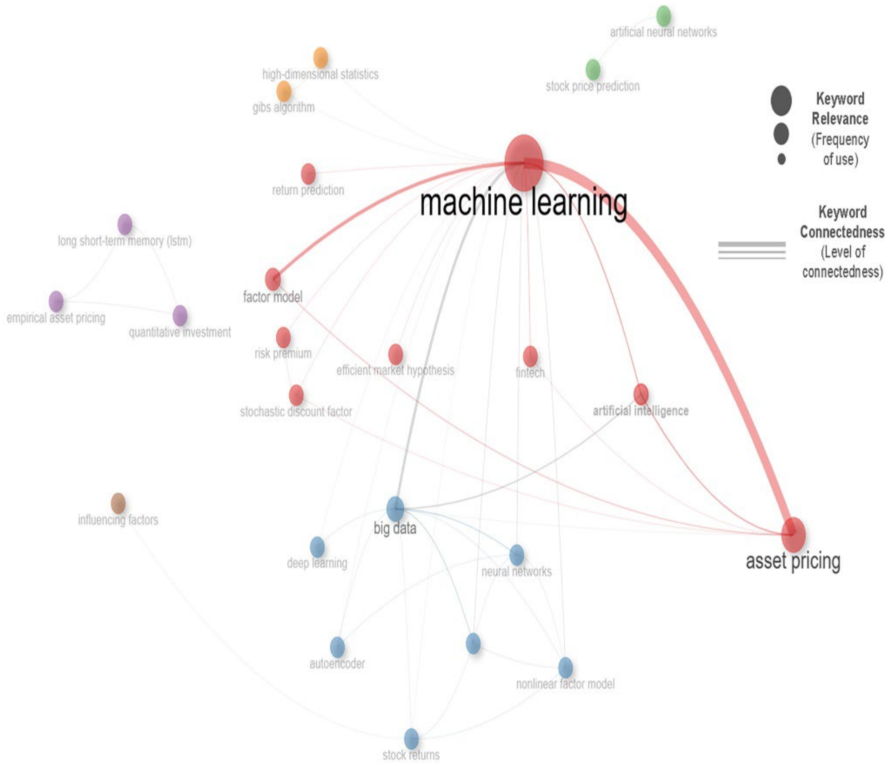



**Figure S1**: Keyword Co-occurrence of AI and Asset Pricing
*Note*: The figure represents how often author(s) keywords appear together in the literature. The circles are keywords, edges show the level of keywords connectedness, and the colour codes are clusters, representing a theme. The thickness of the edges show the strength in connectedness and the size of circles determine frequency/relevance of keywords.

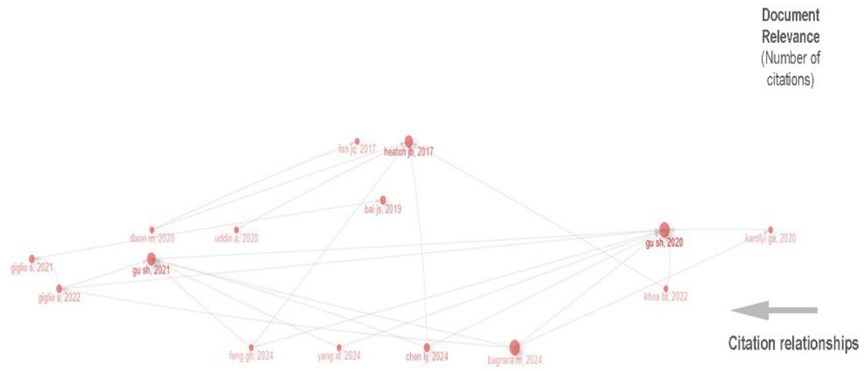



**Figure S2**: Historiograph of AI and Asset Pricing
*Note*: This figure is a historical map of the evolution of literature in AI and asset pricing corpus by showing citation relationships. The number of citations is set at 20 (default in Biblioshiny)

**Table S5**: Bibliographic Coupling and Co-citation Analysis of Big Data and Asset Pricing

**Panel A: Bibliographic Coupling**

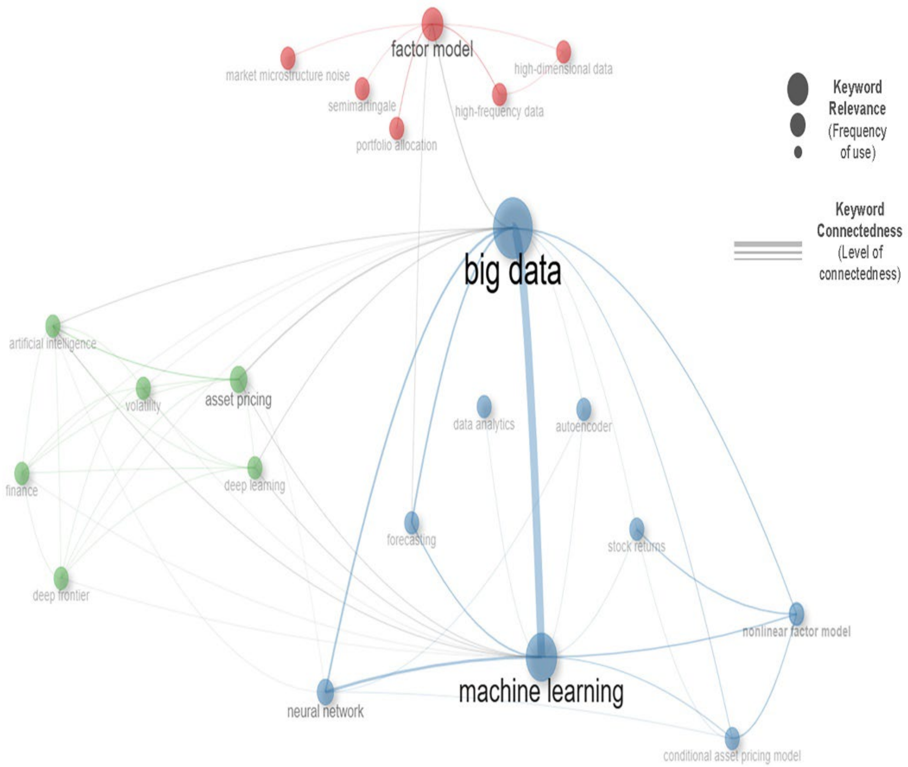
| **Cluster (Theme)** | **Reference** |
|---|---|
| 1<br>(High-frequency and high-dimensional data modelling) | • Dai et al. (2019)<br>• Fan et al. (2016)<br>• He et al. (2022)<br>• Kim and Fan (2019)<br>• Kong et al. (2018)<br>• Lee and Mykland (2012)<br>• Masoliver et al. (2006)<br>• Shephard and Xiu (2017) |
| 2<br>(Improved forecasting accuracy) | • L. Y. Chen et al. (2024)<br>• Costa et al. (2021)<br>• Coulombe et al. (2022)<br>• Gu et al. (2021)<br>• Heaton et al. (2017)<br>• Li and Mei (2020) |
| 3<br>(Advanced asset pricing models for high-frequency and high-dimensional data analysis) | • Aït-Sahalia et al. (2020)<br>• Hollstein et al. (2020)<br>• Ma, Leong et al. (2023)<br>• X. Zhang et al. (2023) |
| 4<br>(High-dimensional factor models) | • Bai and Ng (2002)<br>• Bai (2003)<br>• Shang et al. (2019) |

**Panel B: Co-Citation Analysis**

| **Cluster (Theme)** | **Cited Refence** | **Total Citations** |
|---|---|---|
| 1<br>(Factor models) | • Chamberlain, G., & Rothschild, M. (1982). Arbitrage, factor structure, and mean-variance analysis on large asset markets. | 11 |
| | • Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56. | 15 |
| | • Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1), 1-22. | 10 |
| | • Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273. | 13 |
| | • Ross, S. A. (2013). The arbitrage theory of capital asset pricing. In Handbook of the fundamentals of financial decision making: Part I (pp. 11-30). | 15 |
| | • Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. The journal of finance, 19(3), 425-442. | 12 |
| 2<br>(High-dimensional and high-frequency data and factor analysis) | • Aït-Sahalia, Y., & Xiu, D. (2017). Using principal component analysis to estimate a high dimensional factor model with high-frequency data. *Journal of Econometrics*, *201*(2), 384-399. | 12 |
| | • Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, *71*(1), 135-171. | 11 |
| | • Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191-221. | 15 |
| | • Barndorff-Nielsen, O. E., & Shephard, N. (2004). Econometric analysis of realized covariation: High-frequency based covariance, regression, and correlation in financial economics. *Econometrica*, *72*(3), 885-925. | 11 |
| | • Fan, J., Furger, A., & Xiu, D. (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics*, *34*(4), 489-503. | 12 |
| | • Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167-1179. | 12 |

*Note: a.) Bibliographic coupling connects papers that share (same) references by grouping them into clusters. Shows papers that cite the same references (papers are used as unit of analysis). Thus, papers are assumed to be communicating the same idea based on the shared references. It is a backward citation analysis that shows references cited by 2 or more papers in the data corpus (in this case, a minimum of 10 shared references). There are 21 papers that share a minimum of 10 cited references out of the 55 papers.*

*b.) The co-citation analysis links papers that are frequently cited together by other papers. A cluster shows references that are usually cited today, reflecting their influence in the field (to identify key foundational works). The software uses the cited references as unit of analysis. Using a minimum of 10 citations of a cited reference, only 12 cited references meet this threshold out of 1899 cited references.*



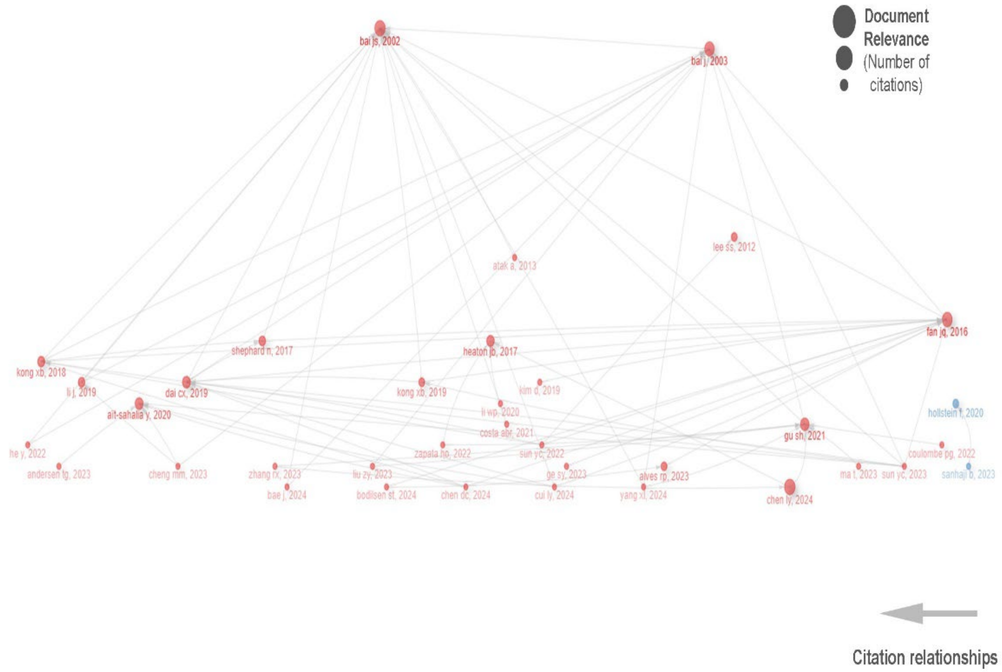




**Figure S3**: Keyword Co-occurrence of Big Data and Asset Pricing

*Note*: The figure represents how often author(s) keywords appear together in the literature. The circles are keywords, edges show the level of keywords connectedness, and the colour codes are clusters, representing a theme. The thickness of the edges show the strength in connectedness and the size of circles determine frequency/relevance of keywords.

**Figure S4**: Historiograph of Big Data and Asset Pricing

*Note*: This figure is a historical map of the evolution of literature in big data and asset pricing corpus by showing citation relationships. The number of citations is set at 20 (default in Biblioshiny).

**Table S6**: Bibliographic Coupling and Co-citation Analysis of Big Data, AI and Asset Pricing

**Panel A: Bibliographic Coupling**

| Cluster (Theme) | Reference |
|---|---|
| 1 (Predictive accuracy of AI models) | • L. Y. Chen et al. (2024)<br>• Gu et al. (2020)<br>• Gu et al. (2021)<br>• Karolyi and Van Nieuwerburgh (2020)<br>• Weigand (2019) |
| 2 (Forecasting abilities of AI models and large datasets) | • Costa et al. (2021)<br>• Coulombe et al. (2022)<br>• Kim and Swanson (2018) |
| 3 (Predictive abilities of AI models) | • Heaton et al. (2017)<br>• Li and Mei (2020) |
| 4 (Advanced models for high-frequency data analysis) | • Centanni and Minozzo (2006)<br>• Shephard and Xiu (2017) |

**Panel B: Co-Citation Analysis**

| Cluster (Theme) | Cited Refence | Total Citations |
|---|---|---|
| 1 (Factor models and AI) | • Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3-56. | 6 |
| | • Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. The Review of Financial Studies, 33(5), 2326-2377. | 5 |
| | • Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273. | 11 |
| | • Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429-450. | 6 |
| | • Kelly, B. T., Pruitt, S., & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501-524. | 5 |
| | • Kingma Jr, J. G., Simard, D., & Rouleau, J. R. (2015). Nitric oxide bioavailability affects cardiovascular regulation dependent on cardiac nerve status. *Autonomic Neuroscience*, *187*, 70-75. | 6 |
| | • Ross, S. A. (2013). The arbitrage theory of capital asset pricing. In Handbook of the fundamentals of financial decision making: Part I (pp. 11-30). | 5 |
| 2 (High-dimensional data analysis) | • Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191-221. | 5 |
| | • Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134-144. | 5 |
| | • Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, *97*(460), 1167-1179. | 5 |

*Note: a.) Bibliographic coupling connects papers that share (same) references by grouping them into clusters. Shows papers that cite the same references (papers are used as unit of analysis). Thus, papers are assumed to be communicating the same idea based on the shared references. It is a backward citation analysis that shows references cited by 2 or more papers in the data corpus (in this case, a minimum of 10 shared references). There are 12 papers that share a minimum of 10 cited references out of the 25 papers.*

*b.) The co-citation analysis links papers that are frequently cited together by other papers. A cluster shows references that are usually cited today, reflecting their influence in the field (to identify key foundational works). Using a minimum of 5 citations of a cited reference, only 10 cited references meet this threshold out of 1196 cited references.*

## Appendix S3: Thematic Analysis

**Table S7**: Identification of Themes

| Theme | Sub-themes | Codes | Reference |
|---|---|---|---|
| | | **Asset Pricing and AI** | |
| Factor Analysis and Identification | Enhance factor analysis and reduced dimensionality | AI models can adequately monitor dynamic factor interactions in portfolios | • Huang (2022) |
| | | Machine learning models enhance reduced factor dimensionality | • Anis and Kwon (2021) • Escribano et al. (2021) • Fan et al. (2017) • Karolyi and Van Nieuwerburgh (2020) • Weigand (2019) |
| | | AI models are effective in enhancing factor analysis | • Fang et al. (2023) • Rui (2023) • Han (2021) • Cui et al. (2024) • Begušić and Kostanjčar (2020) |
| | | Machine learning models enhance forecasting accuracy by leveraging and analysing common and hidden factors | • Maehashi and Shintani (2020) • Chiu and Xu (2004) |
| | Identification of significant and new factors | Identification of new factors from a large pool/factor zoo | • Giglio et al. (2022) |
| | | Identification of significant factors from a large pool/factor zoo | • Bai and Ng (2019) • Weigand (2019) • Cerqueti et al. (2024) • Zhu et al. (2020) • Maasoumi et al. (2024) • Cakici et al. (2023) • Bang et al. (2024) • L. Y. Chen et al. (2024) |
| | | AI models can identify factors in varying market conditions | • Bang et al. (2024) |
| | | AI models/Hybrid models improve prediction/forecasting accuracy | • Fang and Taylor (2021) • Kim et al. (2023) • Escribano et al. (2021) • Jan and Ayub (2019) • Tang et al. (2023) • Khoa and Huynh (2023) • Fang et al. (2023) • Rudkin (2020) • Yao et al. (2024) • Rui (2023) • Khoa and Huynh (2022, 2023) • Luo et al. (2024) • Casarin et al. (2023) • Uddin et al. (2023) • Alaminos et al. (2023) • Levin (2023) • Karolyi and Van Nieuwerburgh (2020) • Lalwani and Meshram (2022) • Lo and Singh (2023) • Huang and Hueng (2008) • Heaton et al. (2017) • Cakici et al. (2023) • Babikir and Mwambi (2017) • Cerqueti et al. (2024) • Yang and Ma (2022) • Yuan and Lee (2015) • Refenes et al. (1994) • Zapata and Mukhopadhyay (2022) • Warin and Stojkov (2021) • French (2017) • Loo (2020) • Healy et al. (2024) • Meng et al. (2024) • Gu et al. (2020) |

| | | | |
|---|---|---|---|
| | | AI models/Hybrid models outperform traditional/linear models used for prediction/forecasting | • Uddin and Yu (2020) • Dixon and Polson (2020) • Anis and Kwon (2021) • Feng et al. (2024) • He and Zhou (2023) • Diallo et al. (2023) • Maehashi and Shintani (2020) • Gu et al. (2021) • Tang et al. (2023) • Zhu et al. (2020) • Wang (2021) • Park et al. (2024) • Zhu et al. (2021) • X. L. Yang et al. (2024) • Khoa and Huynh (2022, 2023) • Jan and Ayub (2019) • De Nard et al. (2022) • Maasoumi et al. (2024) • Pan et al. (2023) • Yao et al. (2022, 2024) • Khoa and Huynh (2022) • Ma, Liao, et al. (2023) • Luo et al. (2024) • J. H. Yang et al. (2024) • Meng et al. (2024) • Uddin et al. (2023) • Weigand (2019) • Kim et al. (2023) • Alaminos et al. (2023) • Huang et al. (2022) • Loo (2020) • Refenes et al. (1994) • Karolyi and Van Nieuwerburgh (2020) • Levin (2023) • Rudkin (2020) • Lo and Singh (2023) • Huang and Hueng (2008) • Ho et al. (2011) • Heaton et al. (2017) • Erfanian et al. (2022) • Khan and Iqbal (2021) • Azevedo et al. (2023) • X. L. Yang et al. (2024) • Lalwani and Meshram (2022) • Chiu and Xu (2004) • Cao et al. (2011) • Yang and Ma (2022) • Cao et al. (2005) |
| | | AI/Hybrid/traditional models that incorporate non-linear parameters have improved forecasting accuracy | • Lo and Singh (2023) • Babikir and Mwambi (2017) • Giglio et al. (2022) • Anis and Kwon (2021) • Maehashi and Shintani (2020) |
| | Improve prediction/forecasting accuracy | AI models/Hybrid models improve the prediction/forecasting accuracy of nonlinear/complex relationships | • Fang and Taylor (2021) • Giglio et al. (2022) • Fan et al. (2017) |
| | | AI models/Hybrid models are effective in capturing complex nonlinear relationships | • Uddin and Yu (2020) • Anis and Kwon (2021) • Escribano et al. (2021) • Feng et al. (2024) • Maehashi and Shintani (2020) • Gu et al. (2021) • Tang et al. (2023) • Zhu et al. (2021) • X. L. Yang et al. (2024) • Pan et al. (2023) • Fan et al. (2017) • Warin and Stojkov (2021) • Lalwani and Meshram (2022) • Alaminos et al. (2023) • Huang et al. (2022) • Karolyi and Van Nieuwerburgh (2020) • Levin (2023) • Babikir and Mwambi (2017) • Bagnara (2024) • Casarin et al. (2023) • Weigand (2019) • Cuomo et al. (2022) • Zapata and Mukhopadhyay (2022) • Begušić and Kostanjčar (2020) • Azevedo et al. (2023) • Gu et al. (2020) |
| Improve Predictive/ Forecasting Accuracy | Capture complex and nonlinear relationships | AI models can offer nonlinear dimensionality reduction | • Anis and Kwon (2021) |
| | | AI models/Hybrid models are effective in accounting for missing data/omitted variables | • Giglio et al. (2021) |
| | | AI models are effective in the presence of large, complex and high-frequency data | • Yao et al. (2022) • Babikir and Mwambi (2017) • French (2017) • Akyildirim et al. (2023) • Cakici et al. (2023) |
| | Robust in handling data | Larger training samples improve AI models performance | • Akyildirim et al. (2023) |

| | | | |
|---|---|---|---|
| | | AI models/Hybrid models are robust for forecasting/prediction and portfolio optimisation | • Dixon and Polson (2020) • Escribano et al. (2021) • Tang et al. (2023) • Wang (2021) • Yao et al. (2024) • Luo et al. (2024) • Giglio et al. (2021) |
| | | Integrating macroeconomic factors into AI/Hybrid models improve model accuracy and ensure robustness | • J. H. Yang et al. (2024) • Erfanian et al. (2022) • Wang (2024) • Bagnara (2024) • Lo and Singh (2023) |
| | Robust models | Machine learning models are effective for regularisation | • Zapata and Mukhopadhyay (2022) • Giglio et al. (2022) • Maasoumi et al. (2024) • Maehashi and Shintani (2020) • French (2017) |
| | | AI models are effective in the long run | • Azevedo et al. (2023) |
| | | The long short-term memory model performs better in long-short strategy performance | • Pan et al. (2023) |
| | | Macroeconomic conditions and firm specific events influence asset prices and shape market movements | • L. Y. Chen et al. (2024) |
| Model Robustness and Adaptability | Adaptability to market conditions | Macroeconomic conditions are important for portfolio/investment decisions | • Wang (2024) • Ho et al. (2011) |
| | | Incorporating a no-arbitrage condition in machine learning asset pricing models ensure robustness and alignment with market principles | • L. Y. Chen et al. (2024) |
| | | AI models are effective in testing asset pricing theories (e.g. EMH, factor models, CAPM) | • O'neill (1995) • Fang et al. (2023) • Caporale et al. (2016) • Weller (2018) • Warin and Stojkov (2021) • Azevedo et al. (2023) |
| | Application in asset pricing theories and anomalies | AI models are effective in exploring asset pricing or market anomalies | • Caporale et al. (2016) • Panchenko et al. (2013) |
| | | AI models uncover better risk-return opportunities by capturing nonlinear relationships between risk factors | • Healy et al. (2024) |
| | | AI/Hybrid models enhance risk management by reducing risk exposure while maintaining market neutrality | • Cuomo et al. (2022) • Weigand (2019) • Meng et al. (2024) |
| | | AI models provide more accurate estimates of volatility and risk pricing | • Levin (2023) • Huang and Hueng (2008) |
| | | AI models enhance predictive accuracy by integrating both historical and current volatility | • Levin (2023) |
| Enhance financial and investment decision making | | AI models/Hybrid models lead to reduced out-of-sample portfolio risk, enhance portfolio performance and return predictions by adapting to varying market conditions | • Auh and Cho (2023) • Gu et al. (2021) • X. L. Yang et al. (2024) • Fang et al. (2023) • Cui et al. (2024) • Ma, Liao, et al. (2023) • Feng et al. (2024) • Diallo et al. (2023) • L. Y. Chen et al. (2024) • Wang (2024) • De Nard et al. (2022) • Duppati et al. (2023) |

| | | |
|---|---|---|
| Improve risk management and portfolio optimisation | | • Begušić and Kostanjčar (2020) • Meng et al. (2024) • Heaton et al. (2017) • Chiu and Xu (2004) • Cuomo et al. (2022) • Park et al. (2024) |
| **Big Data and Asset Pricing** | | |
| Outperforming Hybrid models | Hybrid models consistently outperform traditional asset pricing models and forecasting and prediction models | • Gu et al. (2021) • Li and Mei (2020) • Liu et al. (2023) • Pan et al. (2023) • Shang et al. (2019) • Weigand (2019) • R. X. Zhang et al. (2023) • Chochola et al. (2014) • Dai et al. (2019) • Sun and Xu (2022) • Atak and Kapetanios (2013) • Bodilsen (2024) • Cui et al. (2024) • Cheng et al. (2023) • Martínez et al. (2020) • Avelar and Jordão (2024) • Boubaker et al. (2023) • X. L. Yang et al. (2024) |
| | AI models that are enhanced with higher-order moments outperform traditional models | • H. W. Zhang et al. (2023) • He et al. (2022) |
| | AI models leveraging on big data/high-frequency data outperform traditional models | • Heaton et al. (2017) • Camacho and Lopez-Buenache (2023) • Oomen (2010) • Shephard and Xiu (2017) • D. C. Chen et al. (2024) |
| | Hybrid models improve predictive and forecasting performance | • Heaton et al. (2017) • Zapata and Mukhopadhyay (2022) • Zhan et al. (2022) • Chen et al. (2022) • Kim and Fan (2019) • Bae (2024) • Kiliç et al. (2023) • R. X. Zhang et al. (2023) • Alves et al. (2023) • Liu et al. (2023) • Li and Mei (2020) |
| | Advanced and high-frequency data-driven models outperform traditional methods in managing investments and improving portfolio outcomes | • Fan et al. (2016) • Bodilsen (2024) |
| | Random Forest is superior to other AI models | • Akyildirim et al. (2023) • Costa et al. (2021) • Kiliç et al. (2023) • Avelar and Jordão (2024) |
| | Four-factor IPCA model is superior to the observable factor models and the PCA latent factor models | • Boubaker et al. (2023) |
| | AI models capture complex and nonlinear relationships | • Gu et al. (2021) • Li and Mei (2020) • Weigand (2019) • X. L. Yang et al. (2024) • Masoliver et al. (2006) |
| | AI models efficiently and accurately model nonlinearity | • Gu et al. (2021) • Pan et al. (2023) • Zapata and Mukhopadhyay (2022) |
| | Accounting for higher-order moments in AI models improve the accuracy and effectiveness of predictions | • H. W. Zhang et al. (2023) • Chochola et al. (2014) • Cueto et al. (2020) • X. Zhang et al. (2023) |
| | Models that detect and incorporate jumps for asset pricing improve prediction/forecasting accuracy | • Li et al. (2019) • Lee and Mykland (2012) • Atak and Kapetanios (2013) |
| | Hybrid models are effective in handling large, complex, and incomplete data | • Camacho and Lopez-Buenache (2023) • Dai et al. (2019) |

| | | | |
|---|---|---|---|
| | Model nonlinear and complex relationships | Nonlinear models significantly improve forecasting accuracy | • Coulombe et al. (2022) • Pan et al. (2023) |
| | Robustness and Reliability | Hybrid models significantly improve robustness and reliability in financial modelling | • Dai et al. (2019) • Kim and Fan (2019) • Sun and Xu (2022) • Bodilsen (2024) • Liu et al. (2023) |
| Improve Prediction/ Forecasting accuracy | Leveraging data for improved accuracy | Incorporating relevant data at optimal frequencies enhances asset pricing accuracy and model precision | • Fan et al. (2016) • Oomen (2010) |
| | | Incorporating spatial iteration terms into factor models improves model accuracy | • Ge et al. (2023) |
| | | Using data at its "own" frequency improves forecasting | • Oomen (2010) |
| | | Low frequency data is less accurate for predictions | • Sanhaji and Chevallier (2023) |
| | | High-frequency data provides more accurate and reliable predictions and financial insights. | • Hollstein et al. (2020) • Sanhaji and Chevallier (2023) • Lee and Mykland (2012) • D. C. Chen et al. (2024) |
| | | High-frequency data is useful for capturing time-varying market dynamics | • Bodilsen (2024) • Fan et al. (2016) |
| | | The marginal effect of big data is positive and significant, improving forecasting accuracy | • Coulombe et al. (2022) |
| Optimising Factor Analysis | Factor selection and asset pricing | Enhanced factor models optimise portfolio performance and allocation through advanced factor selection and estimation | • Sun et al. (2023) • Fan et al. (2016) |
| | | Advanced models are effective in identifying, extracting and selecting relevant factors for asset pricing | • Camacho and Lopez-Buenache (2023) • Weigand (2019) • L. Y. Chen et al. (2024) • He et al. (2022) • Sun et al. (2023) • Kong et al. (2019) • Kong and Liu (2018) • Bai and Ng (2002) |
| | | IPCA latent factor models enhance stock return prediction and show significant adaptability to different market conditions | • Ma, Leong, et al. (2023) • Sun and Xu (2022) |
| | | There should be flexibility in factor selection for enhanced model accuracy and adaptation | • Bae (2024) • Bai (2003) • Bai and Ng (2002) |
| | | Imposing a factor structure stabilises risk and uncertainty | • Oomen (2010) • Bai (2003) • Bai and Ng (2002) |
| Portfolio Optimisation and Risk Management | Enhance portfolio efficiency | AI models enhance portfolio efficiency, improve Sharpe ratios and lowers portfolio risk | • Gu et al. (2021) • Cui et al. (2024) • Heaton et al. (2017) |
| | Advanced risk assessment | Advanced models should incorporate jump risk to improve risk assessment and portfolio management | • Li et al. (2019) |
| | | Reduced portfolio risk in longer investment horizons | • Alves et al. (2023) |

| | | | |
|---|---|---|---|
| | | Understanding and accurately modelling volatility is important for predicting returns and market behaviour | • Masoliver et al. (2006) • Peterburgsky (2021) • Alves et al. (2023) |
| | | Hybrid models effectively reduce volatility and lower out-of-sample risks | • Kim and Fan (2019) • Alves et al. (2023) |
| | | Machine learning improves risk premium measurement | • Weigand (2019) |
| | Training data and sample size | Short training models have higher success rates | • Martínez et al. (2020) |
| | | Training periods with larger data sample improve machine learning models | • Akyildirim et al. (2023) |
| | | Enhancing traditional asset pricing models with AI models and/or "rich" (high-frequency/large dimensional) data improves predictive accuracy | • Chen et al. (2022) • Martínez et al. (2020) • Shang et al. (2019) • Aït-Sahalia et al. (2020) |
| Data Quality and Model Training | Leveraging on data quality | High-precision data strengthens market efficiency by minimising risk and uncertainty and enhancing price informativeness and market liquidity | • Zeng et al. (2024) |
| | | To align with market principles and promote robustness, machine learning models should incorporate no-arbitrage conditions | • L. Y. Chen et al. (2024) |
| | | Asset prices exhibit mean reversion and efficient price discovery | • Zhan et al. (2022) |
| | | Returns from different asset pricing models (CAPM, Fama-French) are similar | • Peterburgsky (2021) |
| | Optimising arbitrage and market alignment | Machine learning optimises arbitrage opportunities and strategies | • Zhan et al. (2022) |
| | | Machine learning models are effective across varying time-horizons | • Costa et al. (2021) • Akyildirim et al. (2023) |
| | | AI models are effective for long-term trading and investment strategies | • Pan et al. (2023) • Chen et al. (2022) |
| | Adaptability to time-varying conditions | Advanced asset pricing models are useful in dealing with the dynamics of market microstructure | • D. C. Chen et al. (2024) • Shephard and Xiu (2017) • Peterburgsky (2021) |
| | Influence of macroeconomic and firm-specific events | Macroeconomic conditions and firm specific events influence asset prices and shape market movements | • L. Y. Chen et al. (2024) • Andersen et al. (2023) |

| | | | |
|---|---|---|---|
| Market Dynamics and Macroeconomic Influences | Idiosyncratic factors | News-implied linkages offer a superior method for assessing firm-to-firm connectivity | • Ge et al. (2023) |
| | | Idiosyncratic jumps are related to specific events | • Aït-Sahalia et al. (2020) |
| | | Idiosyncratic betas are time-varying | • Cheng et al. (2023) |
| **Big Data, AI and Asset pricing** | | | |
| Improve Prediction/ Forecasting accuracy | Outperforming AI/Hybrid models | AI/Hybrid models outperform traditional asset pricing and forecasting/prediction models | • Kim and Swanson (2018) • Li and Mei (2020) • Pan et al. (2023) • Weigand (2019) • R. X. Zhang et al. (2023) • Gu et al. (2021) • Liu et al. (2023) • X. L. Yang et al. (2024) • Massa Roldán et al. (2022) • Shephard and Xiu (2017) • Avelar and Jordão (2024) • Gao et al. (2023) |
| | | Machine learning models that account for higher moments outperform traditional models | • H. W. Zhang et al. (2023) |
| | | Models leveraging on complex data patterns outperform traditional benchmark models | • Heaton et al. (2017) |
| | | Random Forest is superior to other AI models | • Akyildirim et al. (2023) • Avelar and Jordão (2024) |
| | | AI/Hybrid models improve forecasting and prediction accuracy | • Alves et al. (2023) • Costa et al. (2021) • Heaton et al. (2017) • Massa Roldán et al. (2022) • Karolyi and Van Nieuwerburgh (2020) • Chen et al. (2022) • Du (2023) • Li and Mei (2020) • Gao et al. (2023) • Zapata and Mukhopadhyay (2022) |
| | | Machine learning models outperform traditional methods for portfolio construction by offering higher risk-adjusted returns/Sharpe ratios | • Gu et al. (2021) • X. L. Yang et al. (2024) • Heaton et al. (2017) • R. X. Zhang et al. (2023) |
| | Model nonlinear and complex relationships | AI models are effective in capturing nonlinear and complex relationships leading to improved forecasting and prediction accuracy | • Coulombe et al. (2022) • Li and Mei (2020) • Weigand (2019) • Gu et al. (2021) • X. L. Yang et al. (2024) • Shephard and Xiu (2017) • Gu et al. (2020) |
| | | Machine learning techniques are efficient for regularisation and modelling nonlinearity | • Zapata and Mukhopadhyay (2022) • Coulombe et al. (2022) • Karolyi and Van Nieuwerburgh (2020) • Pan et al. (2023) |
| | | Incorporating higher-order moments into machine learning models significantly enhances the accuracy of volatility predictions | • H. W. Zhang et al. (2023) |
| | | Linear approximations in nonlinear models can lead to considerable errors in the model predictions | • Gu et al. (2021) |
| | | AI models are robust and reliable for forecasting and prediction | • L. Y. Chen et al. (2024) • Gu et al. (2020) • Liu et al. (2023) |

| | | | |
|---|---|---|---|
| | Model robustness and reliability | Incorporating a no-arbitrage condition in machine learning asset pricing models ensures robustness and alignment with market principles | • L. Y. Chen et al. (2024) |
| Data Quality and Model Training | Training data and sample size | Large training samples improve the performance of AI models | • Avelar and Jordão (2024) • Akyildirim et al. (2023) |
| | Leverage on data quality | The marginal effect of big data improves forecasting/prediction accuracy | • Coulombe et al. (2022) |
| Market and macroeconomic dynamics | Market dynamics | AI models for forecasting perform well in the short-term | • Costa et al. (2021) • Kim and Swanson (2018) |
| | | AI models offer comprehensive analysis and optimisation of various hedging strategies | • Centanni and Minozzo (2006) |
| | | AI models are effective for long-term forecasting | • Kim and Swanson (2018) • Costa et al. (2021) • Chen et al. (2022) • Pan et al. (2023) |
| | Macroeconomic conditions | Macroeconomic conditions significantly influence asset pricing | • L. Y. Chen et al. (2024) |
| Optimise financial models | Advance risk assessment | Reduced realised portfolio risk in longer investment horizons | |
| | | Longer investment horizons and portfolio constraints reduce realised portfolio risk, enhancing stability and risk management | |
| | | AI models deliver a significant reduction in the volatility of minimum variance portfolios | • Alves et al. (2023) |
| | | AI models improve risk premium measurement | • Weigand (2019) |
| | Efficient factor analysis | Machine learning models are effective in selecting most relevant factors/predictors for asset pricing | • Weigand (2019) • L. Y. Chen et al. (2024) |
| | | Machine learning models optimise degree of freedom and reduce factor dimensionality | • Weigand (2019) • Karolyi and Van Nieuwerburgh (2020) |

*Note*: The thematic analysis was guided by Braun and Clarke (2006) 6-phase framework. We used thematic analysis to uncover the patterns in the literature by synthesising the findings to provide a structured overview. The codes are the findings in the data items, which are grouped into related themes (the sub-themes). The related themes are put under an overarching theme for depth and clarity for what has been researched. Hybrid models are conventional asset pricing models that have been framed with AI techniques. Advanced models are traditional asset pricing models that have been enhanced with additional parameters or new models that use traditional asset pricing models as theoretical frameworks.