Tasmanian School of Business and Economics
University of Tasmania

Discussion Paper Series N 2025-03

A Systematic Literature Review of Asset Pricing: Insights from AI and Big Data

Zynobia Barson
University of Tasmania, Australia

Richard Mawulawoe Ahadzie
University of Tasmania, Australia

Dan Daugaard
University of Tasmania, Australia

Joaquin Vespignani
University of Tasmania, Australia

ISBN 978-1-922708-99-1

A Systematic Literature Review of Asset Pricing: Insights from AI and Big Data

Zynobia Barson[1*], Richard Mawulawoe Ahadzie[1], Dan Daugaard[1], Joaquin Vespignani[1,2]

[1]Tasmanian School of Business and Economics, University of Tasmania Hobart, TAS, Australia
[2]Centre for Applied Macroeconomic Analysis, Australian National University Canberra, ACT, Australia

Abstract

This paper systematically reviews the role of big data and artificial intelligence (AI) in asset pricing, analysing 130 journal articles published between 1994 and 2024. We categorise the literature into three themes: AI in asset pricing, big data in asset pricing, and their integration. Publications have grown exponentially since 2019 suggesting structural changes in asset pricing research which we highlight using thematic analysis. The bibliometric analysis shows key trends in AI models for predictive analytics and factor analysis. We identify research gaps and call for adaptive regulatory frameworks to support the ethical use of AI and big data in financial modelling.

Keywords: Big Data, Artificial Intelligence, Asset Pricing, Machine Learning, Systematic Literature Review, Alternative Data, Financial Markets

JEL Codes: G12, C55, C61

* Corresponding author. Email address: zynobia.barson@utas.edu.au

1. Introduction

"Financial markets are among the most dazzling creations of the modern world" (Bernstein, 2012, p. 5). Their constant evolution has led to exponential growth in data, creating the potential to transcend the limitations of traditional asset pricing models. Currently, asset pricing is experiencing transformational shifts through the lens of artificial intelligence (AI) and big data. Big data is an opportunity to move from historical price series to large, varying, and continuous data, thereby providing rich insights into asset prices (Goldstein et al., 2021). In the presence of such high-dimensional conditioning data, AI is useful for accommodating data complexity, ensuring effective analysis and financial modelling (Ahmed et al., 2022; Goldstein et al., 2021; Vespignani & Smyth, 2024).[1] Together, they are transforming asset pricing models into advanced models which enable high-frequency data analysis, enhance predictive performance and, thereby, optimise financial modelling. Amid these excitements, "without a clear picture of where things now stand, simply adding one new study to the existing morass is unlikely to be very useful" (Light & Pillemer, 1984, p. viii).[2] Therefore, we synthesise the asset pricing literature in the era of AI and big data.

Asset pricing models have evolved from the single-factor capital asset pricing model (CAPM) to factor models, and the arbitrage pricing theory (APT) and continue to receive considerable attention particularly, in the era of AI and big data. The literature linking AI, big data and asset pricing is empirically rich but conceptually fragmented with limited synthesis. Thus, to highlight how big data and AI are challenging existing paradigms of asset pricing, this paper seeks to synthesise the literature on asset pricing through a systematic literature review approach. While there are reviews on AI and asset pricing (Ahmed et al., 2022; Bagnara, 2024; Giglio et al., 2022; Goodell et al., 2021; Karolyi & Van Nieuwerburgh, 2020; Meng et al., 2024; Warin & Stojkov, 2021; Weigand, 2019; Zapata & Mukhopadhyay, 2022), there is no

---

[1] Machine learning for algorithmic and high-frequency trading has influenced companies investment decisions as companies are investing heavily into AI-related skills (https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai). PwC projects that AI operations could contribute to up to US$15.7 trillion, a 14% GDP increase in global economy by 2030 (https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf). This projected value creation is supported by forecasts that data colocation centers will generate US$58.4 billion from data storage, cloud computing and AI infrastructure by 2031 (https://explodingtopics.com/blog/big-data-stats). In this *zettabyte* era, big data is offering valuable informational depth for efficient financial decisions. Together, AI and big present an era of advanced real-time data processing and analytics.

[2] Despite the excitement with big data and AI, there are criticisms which represent existential issues. Theoretical criticism lies in the perception that big data is "theory-free" (Coveney et al., 2016), and AI models emphasise statistical patterns in data rather than economic principles (Burrell, 2016). Methodological criticisms persist because AI models are data-driven and are mostly uninterpretable. This presents issues of "quality over quantity", resulting in misleading predictions and spurious relationships, thereby limiting real-world generalisations and economic alignment (Chin-Yee & Upshur, 2019).

synthesised review on big data and asset pricing, and the collective overarching impact of AI and big data on asset pricing. Also, to date, there is no study that has conducted a thematic analysis of the literature on AI, big data, and asset pricing as the extensive reviews on AI and asset pricing predominantly adopt bibliometric analysis. Big data expands the information available for modelling asset prices, while AI introduces new tools for modelling, processing and accurately extracting data relevant for asset pricing. Conducting separate but integrated reviews on big data, AI, and asset pricing allows for a more precise understanding of their independent and collective contributions to reshaping asset pricing. This paper is therefore a synthesis of the literature on big data, AI, and asset pricing, using bibliometric and thematic analyses (see Figure 1).

The findings from the bibliometric analysis reflect a notable geographical concentration of research from China and the United States. Remarkably, AI is used as a model and a tool to enhance prediction/forecasting accuracy, particularly in the context of high-frequency and high-dimensional data. Factor models have also received substantial attention due to the increasing number of factors and the resulting need to identify the more significant factors out of the proliferation of risk factors (hereafter, the "factor zoo"). These findings are further supported by our thematic analysis. We highlight structural shifts in the literature by identifying overarching themes which reveal that the biggest impacts of AI and big data in asset pricing lies in the marginal effect from their complementing characteristics – data processing power and informational depth. The themes reveal improved forecasting/prediction accuracy, optimised financial modelling, usefulness of "rich" and quality data, robustness and adaptability of AI-based models to market and macroeconomic dynamics. This current advancement in asset pricing is a bridge between classical asset pricing and real-world complexities, enabling adaptive asset pricing frameworks, efficient market dynamics, and more informative financial decision making for investors and regulators.

This paper contributes to the transformation in asset pricing literature in three folds. Firstly, we identify themes through a critical appraisal of findings, deepening the knowledge on the impacts of AI and big data on asset pricing. Secondly, we provide insights beyond individual papers on the transformation in asset pricing, by consolidating the literature from both AI and big data simultaneously, thereby extending the systemic review on "*AI and asset pricing*", to "*big data and asset pricing*", and "*big data, AI and asset pricing*". Lastly, we highlight gaps appearing in the literature, serving as a map for future studies to expand the perspectives of empirical asset pricing.

The remainder of the paper is organised as follows. Section 2 describes the data synthesising and collection process. Section 3 is the descriptives, methodological review and a summary of the data scope of the included papers. Section 4 is the findings and discussions; Section 5 suggests future research directions and Section 6 is the conclusion.

2. Data Collection Process and Synthesising

We synthesise the literature on AI, big data and asset pricing, by conducting thematic and bibliometric analyses through a focused and transparent approach of selecting papers – a systematic literature review (Daugaard, 2020; Linnenluecke et al., 2020).

2.1 Justification for Keywords and Search Strategy

The keywords for the main concepts describe their relevance in asset pricing (see Table 1). Understanding the dynamics of asset price discovery lies in appreciating that data process is continuous and time varying, and transactions occur at high-frequency and real time (Goodhart & O'Hara, 1997; Hasbrouck, 2004). Thus, to complement the characteristics of a practical financial data, the keywords for big data reflect the statistical characteristics and real information captured in asset prices (see Engle, 2000; Goodhart & O'Hara, 1997). Given that AI-related studies are extensive, the keywords choice are not restrictive but are prevalent in asset pricing for enhancing predictions, identifying hidden patterns, and exploring interdependencies that tradition models might miss (Bartram et al., 2020; Yang et al., 2020). Ultimately, the keywords for asset pricing are motivated by the seminal works of Carhart (1997), Fama (1965, 1970), Fama and French (1992, 1993, 2004, 2015), Ross (2013) and Sharpe (1964) which continue to serve as the theoretical frameworks for asset pricing models.

We conducted a pilot screening in multiple databases (Business Source Ultimate, IEEE Xplore, Scopus, and WoS) to test the validity of the keywords and saw that a substantial proportion of the papers were captured by the indexing in WoS. Consequently, the search was conducted in "All Databases" collection in WoS. WoS provides access to a comprehensive collection of high quality papers from across all indexed resources since 1637-present (Daugaard, 2020; Daugaard et al., 2024). In Table 1, the search strings were restricted to Business, Business Finance, and Economics research areas, under the Topic field (searches in title, abstract, and indexing).

2.2 Inclusion and Exclusion Criteria

We conducted a rigorous triangulation process to ensure that all the journal publications included in this paper are relevant to the concepts and keywords associated with each research question. Three researchers carried out a manual review – for each question, two researchers contrasted their selection. Where there were conflicts, researchers had to review the document till an agreement was reached. The triangulation process was guided by the inclusion/exclusion criteria (see Table 2) and authors' interpretive judgements, grounded in theoretical alignment and methodological consistency with asset pricing models.

To ensure documents included have undergone a rigorous peer review process, and avoiding literal and machine translation biases, only *journal publications in English* are included. Retracted publications, likely withdrawn due to errors or unreliable findings, dissertations/theses which are subject to personal biases and varied quality control, and books and book chapters consequently having a broader focus and lack rigorous peer review are excluded. Conference proceedings and editorials are also excluded because they are mostly incomplete research pending improvements based on feedback/comments and opinion pieces. Further, we included papers valuing assets in line with asset pricing theories without a restriction to asset types. Papers forecasting returns with an implied or explicit focus on risk premia were also included on a case-by-case basis, where theoretical coherence with the asset pricing literature was evident. Papers based solely on time series forecasting, those on bonds – given how they are valued, and papers forecasting macroeconomic variables unrelated to assets prices were excluded.

2.3 Data Collection

The data collection and screening process was guided by the standards of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 workflow to ensure the trustworthiness and replication of this review process (Page et al., 2021). Using the search strings in WoS, the documents identified are triangulated based on the inclusion and exclusion criteria and authors' interpretive judgements. A two-stage-screening process was implemented to identify eligible papers – first, reading the title and abstract; and secondly, a full text reading to remove "unrelated/unfocused" documents. After screening, 71 papers (for AI and asset pricing), 55 papers (for big data and asset pricing) and 22 papers (for big data, AI and asset pricing) were retrieved for an iteration process.

Through an iteration process ensuring keyword robustness, the search string for AI is updated with "*neural network**" guided by the 10 most relevant author(s) keywords (see Table 3). This is justified because deep learning and some machine learning models are built on neural networks making the keyword collective to AI models (Warin & Stojkov, 2021). With the updated search strings, another search is conducted in WoS for "asset pricing and AI" and "big data, AI and asset pricing" *only*, by adding a date range (*publication date*, customised to the initial search date – September 17, 2024). Applying the inclusion/exclusion criteria to the new documents identified, they are screened for eligibility (in Figure 2, "identification of studies via iteration"). The final papers included in this review are 81 (for AI and asset pricing), 53 (for big data and asset pricing) and 24 (for big data, AI and asset pricing).[3] A list of references of the included papers are in Supplementary Appendix C.

2.4 Data Synthesising

To synthesise the literature on asset pricing, we use bibliometric and thematic analyses. A bibliometric analysis is conducted to identify the trend and evolution in asset pricing literature from the impact of AI and big data using VosViewer and Biblioshiny. Mainly, we present results on citation analysis (bibliographic coupling and co-citation analysis), keyword co-occurrence and historiographs (see Table 4). Bibliographic coupling represents a data corpus most accurately and groups documents with the assumption that they share similar ideas based on shared references (Boyack & Klavans, 2010; Maseda et al., 2022). The limitation of the bibliographic coupling lies in the fact that it may miss related documents because they do not share same references. To complement this limitation, a co-citation analysis which groups documents based on the frequency of references cited together (Donthu et al., 2021), is used to identify the seminal papers that the literature is built on (Boyack & Klavans, 2010; Walter & Ribière, 2013). In a historical context, the foundational papers that have motivated the evolution of knowledge in the asset pricing literature from the impact of AI and big data are presented in historiographs (see Tucker, 2004). Additionally, a keyword co-occurrence map is presented to visualise the literature (see Lee & Su, 2010).

We use thematic analysis to identify meaningful patterns from the included papers to make sense of the transformation in the asset pricing literature. To establish a robust approach to identifying the themes, we adopt Braun and Clarke's (2006) 6-phase framework. This framework provides the foundation to identifying themes which are strongly linked to the theoretical interest of the respective research questions (Braun & Clarke, 2006; Dawadi, 2020). Through a deductive approach, the findings of the literature are systematically coded into common patterns (codes), reviewed, and collated into sub-themes to capture the overarching themes in the literature.

3. Description of Included Papers

The descriptive statistics of the included papers for each theme is described in Table 5. The publication period for AI and asset pricing is 1994:2024 with an annual growth rate of 9.45%, and an average of 18.49 citations per paper. There are 13 single-authored and 68 author-

[3] There are some overlapping papers that are captured across the respective themes. There are 9 overlapping papers in "asset pricing and AI" and "big data and asset pricing"; 9 overlapping papers in "asset pricing and AI" and "big data, AI and asset pricing"; and 18 overlapping papers in "big data and asset pricing" and "big data, AI and asset pricing".

collaborated publications. Publications in "big data and asset pricing" and "big data, AI and asset pricing" date from 2006. For big data, AI and asset pricing, there are 2 single-authored publications and 22 author-collaborated publications, annual growth rate of 6.29% publications, and an average citation of 31.04(5.582) per paper(year). Evidently, in the last 5 years, there has been a surge in publications; indicating the growing interest in big data, AI and asset pricing particularly, in 2023. There are 11 most cited papers in the "big data and asset pricing" corpus – Costa et al. (2021) and Kim and Fan (2019) have 25 citations each. In this corpus, there are 5 single-authored and 48 author-collaborated publications, and an annual growth rate of 12.25% publications. The average citations per paper(year) is 20.75(3.455).

In Figure 3, the annual total publications and citations are presented for each corpus. Generally, there is a surge in total publications from 2017 indicating the gradual integration of AI and big data in empirical asset pricing. Even though AI has been in existence since the 1950s (see Samuel, 1959), this new generation of AI (machine learning and deep learning) is profound in asset pricing because of its data processing power. Empirically, AI models are efficient in learning from data and extract significant information from large datasets that are complex and nonlinear, noisy and evolving (e.g., Bodilsen, 2022; Cao, 2020). This indicates the significant traction in the asset pricing literature to AI and big data for optimised financial modelling.

3.1 Methodological Review and Data Scope of Included Papers

We conduct a methodological review to identify how AI techniques are being applied in asset pricing (see Supplementary Appendix A, Table A1). From the review, AI models are effective in handling high-dimensional data, interpreting temporal signals, and for extracting data from alternative and high-frequency data. They are also efficient for minimising noise and errors and prevent issues of overfitting. These functionalities make AI models robust and adaptive for insightful financial modelling:

- *Regularisation*
- *Dimension reduction*
- *Noise reduction*
- *Feature extraction*
- *Embedded learning*
- *Generative learning*
- *Deep learning*
- *Image-based learning*
- *Ensemble modelling*
- *Interaction modelling*
- *Error minimisation*
- *Nonlinear function approximation*
- *Data classification and clustering*
- *Temporal deep learning*

Researchers have leveraged these functionalities by adopting AI in several ways – as a tool and a model. AI has been used to enhance the predictive accuracy of traditional asset pricing models. Additionally, AI has been used to build hybrid models based on classical asset pricing frameworks (hereafter, "AI-based models"). This consists of a range of *machine learning* and *deep learning models* and *factor models*. Empirically, AI-based models are recognised as a major advancement over traditional asset pricing models.

The literature is largely focused on stock market data from United States and China (see Supplementary Appendix A, Table A2). There are a couple of studies on cryptocurrency, ETFs, commodities and a few on global markets and country-specific data (India, Japan, Korea, Mexico, Pakistan, South Africa, Taiwan, Turkey, Vietnam and United Kingdom). The data are collected and analysed at varied frequencies (second, minute, daily, weekly, monthly and annual data) as influenced by study objective(s).

4. Findings and Discussions

The findings are structured around the trend and evolution, and themes in the asset pricing literature through the lens of AI and big data.

4.1 Bibliometric Analysis

The groups/clusters in the bibliographic coupling and co-citation analysis represent a common theme. In each cluster, the themes are identified from the title and abstract of the respective papers. A thorough analysis of the papers and the objectives are used to draw an overview of the literature in each cluster (see Maseda et al., 2022). In the keyword co-occurrence map, the circles(nodes) are the keywords, and the size determines the keyword(s)' frequency/relevance in the literature. The lines(edges) connect the keywords. The level of connectedness depends on the thickness of the lines. Highly(less) connected keywords are the core(niche) keyword. The colour codes represent a thematic area. A breakdown of the bibliographic coupling (Panel A) and co-citation analysis (Panel B) are in Supplementary Appendix B, Table B2, B3 and B4. Descriptive representations of the bibliographic analysis are in, Table B1, Figure B1, B2 and B3.

4.1.1 Bibliometric Analysis for AI and Asset Pricing

Bibliographic Coupling

In Figure 4, there are 5 clusters from 23 papers that share a minimum of 10 similar references out of 81 papers. The common themes represent the usefulness of machine learning models in enhancing factor models, testing asset pricing theories and handling the dynamics of financial markets.

In the red cluster, the papers contribute to using machine learning models for factor analysis and modelling the challenges in high-dimensional data. The main objective of these papers is using machine learning models to approximate factor models (Fan et al., 2017; Bai & Ng, 2019; Giglio et al., 2021, 2022); methodological advancements in factor models (Gu et al., 2021; Uddin & Yu, 2020) and comparative analysis of AI-based models and traditional factor models (Maehashi & Shintani, 2020). The general overview of these papers is the advancement in factor models for asset pricing by capitalising on the functionalities of machine learning.

The literature in the green cluster communicates how AI-based models accommodate high dimensional and complex data and are useful in testing theories of asset pricing. Caporale et al. (2016) show how statistical anomalies do not attribute to market efficiency based on a trading robot analysis by incorporating transactions costs. Huang and Heung (2008) investigate the asymmetric risk-return relationship in a time-varying beta CAPM and find that returns depend on the level of risk assumed. Weller (2018) analyses how information acquired through algorithmic trading may be incorporated into asset prices and finds that despite the importance of algorithmic trading in this process, it may reduce the effectiveness of price discovery. In testing traditional models, Refenes et al. (1994) used neural networks as a forecasting technique in the framework of APT which outperforms linear regression models. Subsequently, in the advancing literature, Yuan and Lee (2015) and Ho et al. (2011) develop an augmented forecasting model to predict systematic risk and for enhanced portfolio selection based on CAPM. Erfanian et al. (2022) also investigates the influence of economic theories on Bitcoin prices and find that traditional models may be more effective in predicting the prices of Bitcoin than machine learning models. The overview of these papers is a growing interest in using AI models to enhance the accuracy of traditional asset pricing.

The blue cluster represents the integration of deep neural networks and traditional asset pricing models for theoretical and empirical applications. These papers emphasise the superiority of AI-based models to traditional asset pricing models (Cao et al., 2005, 2011). L. Y. Chen et al. (2024) and Heaton et al. (2017) also show how deep learning models can be used for predictions and optimise portfolio construction. The overview of these papers highlights the growing role of AI models in enhancing traditional asset pricing models. This is captured by Weigand (2019)'s empirical review of machine learning in asset pricing.

The yellow cluster represents how machine learning models are improving factor models (Jan & Ayub, 2019), by reducing factor dimensionality and identifying relevant factors from the expansive factor zoo (Karolyi & Van Nieuwerburgh, 2020). Generally, this cluster represents studies contributing to the methodological advancements in factor analysis using AI models.

The purple cluster captures literature contributing to theoretical exploration of market dynamics and methodological advancements in asset pricing. Warin and Stokjov (2021) conduct a metadata-based review of neural networks and machine learning models in asset pricing. Panchenko et al. (2013) investigates the effects of network topologies on the dynamics of asset pricing. These papers represent an understanding of the application of machine learning in financial analysis.

Co-Citation Analysis

The co-cited references in this data corpus are foundational works in factor models and asset pricing theories (see Figure 5). These co-cited references are relevant for understanding some of the foundational works on factor models and the shift from traditional asset pricing models to AI models.

The red cluster represents seminal literature on factor models and their era-defining contributions. Cochrane (2011) and Fama and MacBeth (1973) lay some of the groundwork for factor models by testing the traditional view of asset pricing. They introduce new factors that can explain price dividend ratios, test the relationship between average return and risk, and how common market factors affect the performance of mutual funds (see also, Carhart, 1997). The literature has since moved to exploring large pool of factors by addressing the issue of which factors are relevant for asset pricing (Feng et al., 2020), developing advanced factor models (Hou et al., 2015), and setting the significant threshold for identifying new factors (Harvey et al., 2016). Kozak et al. (2020) also demonstrate that most of the new factors are redundant to existing factors and further contribute to improving the stability and interpretability of factor models by shrinking the cross-section of asset returns. Kelly et al. (2019) identify latent factors using a unified model of risk and return while Lettau and Pelger (2020) identify factors that fit both time series and cross-section of stock returns. To challenge traditional asset pricing models, Freyberger et al. (2020) explore the importance of including non-linear parameters for robust financial modelling. Gu et al. (2020), Gu et al. (2021) and Heaton et al. (2017) are useful to understanding the adoption of AI in asset pricing. Heaton et al. (2017) detect complex interactions for accurate predictions using deep learning which is complemented by Gu et al. (2020, 2021). Gu et al. (2021) demonstrates how machine learning models are useful for capturing non-linear and latent factor exposures and for identifying significant economic gains for investment decisions (Gu et al., 2020).

The seminal papers on asset pricing theories are commonly cited together in the green cluster. These papers are relevant to understanding the theoretical and empirical evolution in asset pricing (Fama & French, 1992, 1993, 2015; Lintner, 1975; Sharpe, 1964; Ross, 2013). Sharpe (1964) draws a theory on the relationship between risk and return, highlighting the importance of market risk in asset pricing (CAPM) which is extended by Litner (1965), emphasising the role of risk in asset valuation. Fama and French (1992, 1993, 2015) extend the

CAPM to factors beyond market risk to explain the cross-sectional differences in asset prices in their 2-,3- and 5-factor models. To avoid some of the restrictive assumptions in the factor models, the work of Bai and Ng (2002) is foundational in specifying the number of factors accurate for consistent estimations. Ross (2013) also proposes a flexible framework that accommodates multiple factors for asset pricing. These papers provide a strong foundation for understanding the dimensions of factor analysis by highlighting the different factors that explain asset prices.

Keyword Co-occurrence

In Figure 6, there are 6 clusters. In the red cluster, there is evidence of increasing interest in the integration of AI models to traditional asset pricing concepts. Machine learning is most dominant in the AI-asset-pricing literature because both AI and machine learning are conceptually interrelatedly (Goodell et al., 2021). Machine learning has a higher level of connectedness to factor model, big data and asset pricing. Further, the keywords in the blue cluster are a representation of the usefulness of AI in handling data complexity and extracting significant, hidden and new information from big data. The implication of this keyword co-occurrence is the shift to AI-driven models for asset pricing.

Historiograph

In Figure 7, there are two dominant papers that are mostly cited: Heaton et al. (2017) and Gu et al., 2021). The trajectory of literature from Heaton et al. (2017) reflects the integration of traditional asset pricing models with AI to handle data complexity and identify hidden and short-lived patterns. Khoa and Huynh (2022) compare the predictive accuracy of traditional models to machine learning within a multiple factor model framework (5-Fama-French factor model) and show that AI improves the accuracy of factor models and effectively exploits latent factors. Uddin and Yu (2020) also develop a dynamic network structure to capture market conditions and extract latent factors that can explain and predict risk premia. To further show the effectiveness of AI models in data processing, Dixon and Polson (2020) use a deep fundamental factor model to capture non-linearity and data interactions showing the model's superiority to ordinary least square and least absolute shrinkage and selection operator models. L.Y. Chen et al. (2024) also adopts large conditioning information for asset pricing by using deep learning while maintaining flexibility for varying time conditions.

Likewise, Feng et al. (2024) generate latent factors for cross-sectional asset pricing through a deep factor model and find the model useful for capturing non-linearity. Baganara (2024) and Giglio et al. (2022) are review papers on machine learning' s contribution to asset pricing. Bagnara (2024) highlights the functionality of AI models in asset pricing and Giglio et al. (2022) collate the methodological contributions of AI in factor models in the presence of factor zoos. Their contribution to the literature from Gu et al. (2021) highlight the dominance of AI models over traditional asset pricing models by identifying the methodological contributions and functionalities of AI models in asset pricing. This is further advanced by L. Y. Chen et al. (2024), Feng et al. (2024) and Yang et al. (2024) who provide evidence for the effectiveness of AI in capturing complexity in data.

4.1.2 Bibliometric Analysis for Big Data and Asset Pricing

Bibliographic Coupling

In Figure 8, 18 papers (out of 53 papers) are coupled into 4 clusters. The clusters invariably show the application of machine learning and deep learning models to asset pricing and for advanced forecasting analysis by leveraging on high-frequency data. Thematically, this literature shows a shift to high-dimensional and high-frequency data for asset pricing, forecasting and predictions based on its informational depth.

There are 6 papers in the red cluster which explore the usefulness of machine learning and deep learning for asset pricing and forecasting. The objectives of these papers are inclined to identifying how AI models promote accurate and adaptable asset pricing models (L. Y. Chen et al., 2024: Costa et al., 2021; Heaton et al., 2017; Coulombe et al., 2022; Gu et al., 2021). An empirical review from Weigand (2019) show a large shift to machine learning techniques for asset pricing based on its data processing ability. Largely, the overview of this cluster strongly emphasises on the improved accuracy of forecasting analysis and how advanced models are useful for asset pricing.

The green cluster represents literature on high-frequency and high-dimensional data in addressing the challenges of asynchronicity, jumps and noise. Dai et al. (2019) explore different approaches to analysing large covariance matrices and constructs a pre-averaging-based large covariance matrix estimator which is robust to noise. Fan et al. (2016) also propose a matrix estimator for high-frequency data while Kong et al. (2018) test for the consistency in the loading matrices of high-frequency data. Lee and Mykland (2012) differentiate between the large shifts in equilibrium prices from market microstructure noise by proposing a nonparametric test which asymptotically removes noise from data. Shephard and Xiu (2017) also develop a multivariate realised quasi maximum likelihood approach which achieves an efficient rate of convergence for estimating the dependence in financial assets. The literature in this cluster is a clear shift towards using high-frequency data for enhanced accuracy in financial modelling and analysis.

Literature focused on using high-frequency data for enhanced and accurate financial modelling are in the blue cluster. Aït-Sahalia et al. (2020) develop a non-parametric regression model to estimate betas; Hollstein et al. (2020) tested conditional CAPM using high-frequency data to see if their model can better explain asset pricing anomalies; and H. W. Zhang et al. (2023) explore the predictive ability of higher-order moments using high-frequency data.

The dominance of the literature in the yellow cluster is factor analysis (He et al., 2022; Kim & Fan, 2019) and how volatility in financial markets influence asset prices (Masoliver et al., 2006). Masoliver et al. (2006) propose a framework that is suitable for modelling asset price dynamics and volatility by considering the frequency level changes in financial markets. He et al. (2022) develop a factor model that is robust for heavy-tail large dimensional data analysis and Kim and Fan (2019) propose a dynamic factor model with high-frequency stochastic processes to predict volatility. This cluster provides an overview of the usefulness of high-frequency and high-dimensional data in factor analysis and volatility predictions.

Co-Citation Analysis

The analysis of these references reveals the evolution of the asset pricing literature to advanced asset pricing models (especially for factor analysis), and the integration of machine learning models for accurate financial analysis (see Figure 9). The co-cited references show a convergence of the literature to factor models and high-frequency data which is not surprising as the dimensionality of information in financial markets keep increasing.

The red cluster captures literature on asset pricing theories and factor-based models. Sharpe (1964) and Ross (2013) are known for their groundbreaking works in asset pricing by explaining asset returns based on market risk and multiple risk factors. Additionally, Fama and French (1993) highlight how a 3-factor model can better explain asset returns as compared to CAPM. Chamberlain and Rothschild (1982) also communicate how arbitrage can exist in markets by conceptualising approximate factor structures, implying that asset returns can be explained by only a marginal number of factors. Intuitively, the papers in this cluster contribute to the foundation of multiple factor models. Gu et al. (2021) further advance on using machine learning models for asset pricing as it adequately captures non-linear and high-dimensional data to improve on forecasting accuracy as compared to traditional linear factor models.

The papers in the green cluster explore the usefulness of high-frequency data. Inherently, they show that high-frequency data can enhance factor-based models and is crucial for understanding the dynamics of market liquidity. Aït-Sahalia et al. (2020) draw on the conditioning information from high-frequency data to identify new factors that explain the empirical properties of stocks from the NYSE, AMEX and NASDAQ. Likewise, Fan et al.

(2016) integrates a comprehensively high-dimensional data into covariance estimations and show improved portfolio estimation strategies. Collectively, the papers in this cluster show the usefulness of using high-frequency data in improving financial modelling based on the depth of conditioning information.

The blue cluster represents advancements in factor analysis. It communicates the usefulness of high-dimensional data in identifying significant factors driving asset prices. Bai (2003) develop an inferential theory for factor modelling of large dimensional data by focusing on principal components estimators, allowing for efficient identification and measurement of factors driving asset prices. Based on the increasing pool of factors, Bai and Ng (2002) also develop a statistical measure to determine approximate factors for effective factor analysis by reducing dimensionality just as Stock and Watson (2002), who explore the use of PCA in the presence of large factors. Largely, these papers propose the usefulness of advanced models for reducing factor dimensionality.

Keyword Co-occurrence

In Figure 10, the level of connectedness between big data and factor model confirms the existence of increasing factors. This connectedness is relatively low as compared to the connectedness in the most frequent co-wording (big data and machine learning). In the blue cluster, big data is highly connected to machine learning and exhibits relative connectedness to neural network and autoencoders. This indicates the complementing characteristics of AI models and big data. Graphically, the connectedness is stronger between the green (asset pricing) and blue (machine learning) clusters. The green cluster depicts the application of AI in asset pricing and for modelling volatility. In the red cluster, the keyword empirically, represent the characteristics of factor models (high-frequency and high-dimensional data) and how factors guide portfolio construction. This multi-connectedness reveals how machine learning may be useful for handling challenges in factor zoos.

Historiograph

The historiograph of big data and asset pricing are highly connected yet, Gu et al. (2021) and Fan et al. (2016) are the most cited (see Figure 11). Papers that cite Gu et al. (2021) contribute to the literature on the effectiveness of AI models in handling complexities and uncovering data patterns (Coulombe et al., 2022; Zapata & Mukhopadhyah, 2022) and, in enhancing predictive analysis (L. Y. Chen et al., 2024; Liu et al., 2023; R. X. Zhang et al., 2023). Intuitively, the papers in this cluster (i.e., citing Gu et al. 2021) explore the predictive accuracy of AI models in the presence of high-dimensional data. The most connected cluster (red), mostly cite Fan et al. (2016), and contribute to the literature on high-frequency data and advanced factor models.

Fan et al. (2016) highly connects to the works of Kim and Fan (2019) and Sun et al. (2022, 2023), Kong et al. (2019), Bodilsen (2022) and Alves et al. (2024) – who use high-frequency-based factor models to predict covariance matrices. Capitalising on high dimensional data, D. C. Chen et al. (2024), Cui et al. (2024) and Shephard and Xiu (2017) propose advanced models that can handle market microstructure noise and weak factors for enhanced predictions. The influence of Fan et al. (2016) in developing a model to forecast time series when there is high-dimensional data further extends to the works of Dai et al. (2019) and Ge et al. (2023) who explore effective portfolio allocation and risk management strategies using advanced models. The evolution of literature from Fan et al. (2016) is the use of high-frequency and high-dimensional data for improved accuracy and reliability of financial models.

4.1.3 Bibliometric Analysis for Big Data, AI and Asset Pricing

Bibliographic Coupling

There are 3 clusters of 9 papers sharing a minimum of 10 cited references out of the 24 papers in this data corpus (see Figure 12). These clusters inform the advancement of the literature to AI-based models for asset pricing, and their usefulness in handling the complexity of high-frequency and high-dimensional data.

The red cluster reflects how machine learning and deep learning models are used for efficient asset pricing and improve forecasting/prediction accuracy. Of the 5 papers in this cluster, Karolyi and Van Nieuwerburgh (2020) and Weigand (2019) are reviews informing the shift to machine learning models for asset pricing and the benefits of reduced dimensionality and identification of new factors out of the zoo. L. Y. Chen et al. (2024) and Heaton et al. (2017) also show how deep learning models are applicable to asset pricing and for predictions. Gu et al. (2021) develop an asset pricing model using deep learning for enhanced accuracy.

The green cluster highlights the role of machine learning for enhanced forecasts and improved reliability. Kim and Swanson (2018) sought to explore which advanced method could improve economic forecasts; Coulombe et al. (2022) explored the characteristics of machine learning models that improve macroeconomic forecasts and Costa et al. (2021) use machine learning for improved accuracy by capitalising on high-dimensional data. These papers provide evidence on the usefulness of big data in machine learning models for enhanced forecasting accuracy, highlighting how the models effectively handle large, complex data across different economic conditions.

There are only two papers in the blue cluster, highlighting how advanced models are useful in dealing with the complexity of high-frequency data. Centanni and Minozzo (2006) propose a model for ultra-high-frequency data to predict prices and guide hedging strategies.

Shephard and Xiu (2017) also propose a sound and efficient model for high-frequency data to enhance predictions. These papers show how AI models are useful in reducing noise, capturing useful information such as time-dependencies from high-frequency data.

Co-Citation Analysis

From this data corpus, it is evident that big data is contributing to a shift from traditional asset pricing models to AI-based asset pricing models (see Figure 13). The implications of these co-cited references reflect the need for advanced asset pricing models that can effectively deal with the complexities of high-dimensional and high-frequency data.

The theme of the red cluster is forecasting analysis. The usefulness of multiple factors in asset pricing was explored by Fama and French (1993), introducing the 3-factor model to advocate on the importance of risk factors in asset pricing. Ross (2013) also contributes to the literature by theorising a multi-factor asset pricing model (APT), making it accommodative to market complexities. Bai and Ng (2002) also address the complexity of factor zoos, by inferring the optimal number of factors, just as Stock and Watson (2002). Highlighting the importance of correctly specifying factors in asset pricing, it can be concluded that the literature that have advanced on factor analysis in the event of factor zoos draw motivation from these papers. The implication of these studies has contributed to improved factor analysis in the presence of a continually increasing factors.

The dominance of the green cluster is machine learning in asset pricing. Gu et al. (2020) and Gu et al. (2021) use machine learning to measure risk premia. The results show improved pricing errors when compared to traditional models. Kelly et al. (2019) underscore that factor analysis can be more effective by capitalising on models that can explain a high-dimensional cross-section of returns as compared to existing traditional factor models, highlighting the usefulness of large data sets.

Keyword Co-occurrence

Figure 14 represents the literature on big data, AI and asset pricing. The 2 clusters accentuate the increasing and strong connectedness of machine learning and asset pricing. In the red cluster, AI models (autoencoders, machine learning and neural networks), big data, traditional asset pricing models (CAPM) and nonlinear factor models, exhibit varying degrees of connectedness. The strong connectedness between machine learning and big data highlights their complementary role in overcoming the limitations of traditional asset pricing models. (Goodhart & O'Hara, 1997). The blue cluster represents the transformation in asset pricing due to the integration of AI models. The interconnectedness between the keywords in this cluster, and to machine learning and big data show the growing dominance of AI and big data in asset pricing.

Figure 15 is a keyword timeline displaying an intriguing level of connectedness between traditional asset pricing (CAPM, forecasting, (nonlinear) factor models, portfolio allocation and volatility), AI (autoencoder, deep frontier, deep learning, machine learning, neural network) and big data. This is particularly the case from 2020 to 2023, reflecting the recent evolution to adopting AI and big data for advanced forecasting and factor analysis, and asset pricing. There is also a clear transition of the literature to AI models in most parts after 2020. The dominance of machine learning and big data is also seen in the early part of 2020. From 2020, there is increased connectedness between big data and machine learning in traditional asset pricing models, representing their collective integration.

Historiograph

Gu et al. (2021) is evidently dominant in this data corpus (see Figure 16). Gu et al. (2021) built a latent factor CAPM by leveraging on autoencoder neural networks, to develop a model that reduces factor errors and is flexible for modelling nonlinear relationships in large dataset. Evolving literature include L. Y. Chen et al. (2024), Liu et al. (2023) and X. L. Yang et al. (2024). These studies explore the superior predictive abilities of machine learning models for high-dimensional data, demonstrating its effectiveness in capturing complex relationships while maintaining flexibility for varying time conditions. Through image-based algorithms, R. X. Zhang et al. (2023) use deep learning models to improve price predictions. Coulombe et al. (2022) and Zapata and Mukhopadhyah (2022) also explore the characteristics of machine learning that make them superior to traditional asset pricing models. They identify that their superiority lies in their ability to handle large, model complex relationships and nonlinearity, apply regularisation techniques and support robust cross validation. The literature on asset pricing has since evolved to using machine learning and deep learning models to enhance predictive accuracy.

4.1.4 Discussion of Bibliometric Analysis

Asset pricing is moving away from theory-driven models and is no more limited by data. The integration of AI and big data is reshaping asset pricing, toward more empirically advanced, adaptable and data-intensive approaches. With the extensively available data, AI offers data processing abilities and learning techniques which are effective for modelling and extracting information for insightful asset pricing. The implication of this is efficient financial decisions, improved predictive performance, and optimised risk management and portfolio strategies.

The results from the bibliographic analysis indicate that factor models have received a lot of attention. Theoretically, the extensive interest in enhancing factor analysis is an empirical test of classical asset pricing models (Cahart, 1997; Fama & French, 1993, 2015). The increasing numbers of factors present challenges of dimensionality, model overfitting, and factor redundancy which traditional factor models cannot address. AI models however penalise irrelevant factors through regularisation, identify latent factors that are not immediately apparent, and capture dynamic factors ensuring that only the most predictive and significant factors are retained (e.g., Bai & Ng, 2019; Cakici et al., 2023; Cerqueti et al., 2024; Giglio et al., 2022; Karoyli & Van Nieuwerburgh, 2020; Maasuomi et al., 2024). The main benefit of AI for factor analysis is in its ability to analyse hundreds of factors, reduce dimensionality and improve factor selection. Empirically, AI enhances factor models by making them more flexible, robust, predictive and adaptable to real time factor analysis.

AI models have consistently demonstrated superior predictive accuracy, while offering robust and accurate predictions. A significant contribution for the accuracy of AI models arises from its ability to handle the complexity in high-frequency and high-dimensional data (Camacho & Lopez Buenache, 2023; Martínez et al., 2020; Oomen, 2010; Shang et al., 2019; Yao et al., 2022). AI models are uniquely capable of extracting and processing high-dimensional and granular data to uncover subtle and short-lived patterns. By leveraging the informational depth of high-frequency and high-dimensional data, AI enhances asset pricing by adapting to evolving market conditions in real-time, making the models more responsive and adaptive to dynamic conditions. The integration of high-frequency data allows real-time adaptability for a more comprehensive and dynamic understanding of asset prices. This highlights the complementing features of AI and big data making them useful for asset pricing (Aït-Sahalia et al., 2020; L. Y. Chen et al., 2024; Fan et al. 2016; Gu et al., 2020; Hollstein et al., 2010).

Intuitively, AI does not only allow for augmenting traditional asset pricing models but also opens new avenues for more dynamic, adaptive, and accurate models that better reflect the complexities of modern financial markets (e.g., Centanni & Minozzo, 2006; Coulombe et al., 2022; Costa et al., 2021; Kim & Swanson, 2018). The integration of AI in asset pricing ranges from its ability to model non-linear and complex relationships, flexibility in incorporating macroeconomic variables, to relaxing the assumptions of traditional models, and further enhancing prediction accuracy (Ani & Kwon, 2023; Gu et al., 2021; Li & Mei, 2020). This leads to improved out-of-sample R-squared and Sharpe ratios indicating better economic significance (e.g., Alves et al., 2023; Cui et al., 2024; Kim & Fan, 2019). Unarguably, AI models are useful for asset pricing beyond improving model accuracy, addressing model overfitting and misspecification, discovery of significant factors, and enhancing out-of-sample performance.

4.2 Thematic Analysis

In Appendix A, Table A1, the codes and their respective references, sub-themes and themes are presented. As the analysis progressed, several sub-themes overlapped across the research questions revealing the complementing role of AI and big data in asset pricing. To avoid redundancy, we focus on *only* discussing the themes for how AI and big data are collectively transforming asset pricing. The themes are then discussed in dimensions of sub-themes across the research questions, enabling a more nuanced understanding of the transformation in asset pricing.

First, we identify that AI models *improve the accuracy of predictions and forecasts*, captured in the following sub-themes: outperform traditional models; model nonlinear and complex relationships; and are robust and reliable models. AI models improve forecasting accuracy by efficiently capturing nonlinear and complex relationships in high-dimensional and high-frequency data. Their adaptive learning process facilitates real time data analysis, enabling them to effectively handle complex and nonlinear relationships (e.g., L. Y. Chen et al., 2024; Peterburgsky, 2021). These models effectively account for missing data, can uncover latent patterns and extract meaningful data particularly from large datasets (regardless of structure), without suffering from dimensionality or overfitting issues. As a result, they consistently outperform traditional asset pricing models in predictions (e.g., Alaminos et al., 2023; Bodilsen, 2024; Fang & Taylor, 2021). Additionally, they adjust to stylised facts, market and macroeconomic dynamics which enhances their accuracy as compared to traditional models (D. C. Chen et al., 2024; Kim & Swanson, 2018; Wang, 2024). By capitalising on big data, AI models reduce noise and error while incorporating rich conditioning information, making them robust and reliable tools for financial analysis (L. Y. Chen et al., 2024; Dai et al., 2019; Sun & Xu, 2022).

Secondly, using AI and big data in asset pricing has substantially led to *optimised financial modelling* as in the following sub-themes: factor analysis and identification; improved risk and portfolio optimising strategies; and aligned market and arbitrage opportunities. AI has significantly advanced factor modelling. It has enabled the discovery of new and latent factors, thereby, avoiding factor redundancy despite the existence of new factors almost every day (e.g., Bai & Ng, 2019; Cerqueti et al., 2024; Escribano et al., 2024; Maasoumi et al., 2024; Rui, 2023). Researchers have also advanced on the factor model framework to build AI-based models (e.g., Dixon & Polson, 2020; Pan et al., 2023; X. L. Yang et al., 2024). These advanced models have superior explanatory power, offering a more robust approach for explaining asset returns, thereby aligning markets and arbitrage opportunities (L. Y. Chen et al., 2024; Zhan et al., 2022). The integration of AI and big data into financial analysis contributes to optimised strategies for risk and portfolio management because investors can efficiently capitalise on arbitrage opportunities (e.g., Auh & Cho, 2023; Begušić & Kostanjčar, 2020; Cuomo et al., 2022; Diallo et al., 2023; Duppati et al., 2023). Empirically, AI models uncover better risk-return opportunities by capturing nonlinear relationships between risk factors (Healy et al., 2024; Levin, 2023) and are effective in long-term investment portfolio decisions (Alves et al., 2023).

Evidently, AI-based models are adaptable to market dynamics and efficient when macroeconomic variables are accounted for, hence the theme, *market and macroeconomic dynamics*. Unlike traditional asset pricing models, the literature shows that AI models have flexible parameters for incorporating macroeconomic variables (e.g., Ho et al., 2011; Wang, 2024). Empirically, macroeconomic conditions influence asset pricing, by shaping expectations for returns and provide behavioural and economic justifications for cross-sectional differences in asset prices (Andersen et al., 2023; Aït-Sahalia et al., 2020; Chen et al., 2023). Justifiably, AI-based models incorporating macroeconomic dynamics account for market-wide effects for enhanced accuracy (Ge et al., 2023; Cui et al., 2024). Likewise, AI-based models accounting for time-varying conditions perform better than traditional asset pricing models. Empirical evidence shows that AI-based models successfully identify hidden and temporal dependencies, making them well suited for capturing time-varying parameters for asset pricing (e.g., Akyildirim et al., 2023; Costa et al., 2021; Shephard & Xiu, 2017).

Lastly, AI models trained on data quality and optimal sample size to enhance performance, which we thematise as *data quality and model training*. The literature corroborates on the level of informational depth AI models can capture from high-dimensional and high-frequency data for enhanced accuracy (e.g., Aït-Sahalia et al., 2020; Lee & Mykland 2012; Sanhaji & Chevallier, 2023). Technically, the marginal effect of high-quality data contributes to enhanced model accuracy particularly at optimal data frequencies. The accuracy of AI models however does not depend on the conditioning information extracted from data alone. The frequency at which the data is collected, and the percentage of the data used to train a model contribute to its effectiveness (e.g., Akyildirim et al., 2023; Martínez et al., 2020). Empirical evidence shows that data frequency directly affects model performance, reduce issues of overfitting, and enhance prediction accuracy (Fan et al., 2016; Oomen, 2010).

5. Future Research Directions and Implications

In this new phase of AI and big data contributing to significant advancements in asset pricing, we highlight the following future research areas and their implications (see Figure 17).

5.1 Data

The literature on high-frequency and high-dimensional data is extensive but there is a geographical gap and limited use of alternative data. A geographical gap exists for comparative and country-specific studies. The literature is highly concentrated on economically significant and liquid markets particularly, China, Europe, and United States. With a geographical inclusion, we would better understand the impact of AI and big data on asset pricing, validating their usefulness for asset pricing beyond developed and technologically inclined markets. From the literature, there is yet an overdependence on historical data – as stock market and factor data have received a lot of attention. Outside historical data is an exponential increase in alternative data which could be informative for asset pricing (Goodell et al., 2021).[4] There are other data types that could benefit from AI and big data in its pricing process, such as agricultural, energy, and metal commodities, derivatives, digital assets and real estate markets. Researching the integration of AI and big data across these asset types has significant implications for investors in mitigating risk and for portfolio management decisions.

5.2 Ethics

The use of AI and big data is subject to ethical risk. The extensive availability of data from open sources and connectedness to the internet, make a lot of publicly available data useful for financial analysis. However, open-source AI models may incorporate data from unverified or unauthorised sources, resulting in breach of privacy and confidentiality. Further, as the domain of data collection is widening, it is relevant to predefine "rich and quality" informative data sources. The challenge of choosing the "right" informative data sources and deciding on what sample size to train AI models with is subject to data and model bias. There is also the issue of uninterpretable results from AI models. While the asset pricing literature has begun to address this, their efforts lack consistent standard/benchmark guiding how results from AI-based models should be interpreted (see Dixon & Polson, 2020; Haung, 2022; R. X. Zhang et al., 2023). These unresolved ethical issues highlight the need for a comprehensive examination of the ethical and regulatory considerations associated with the use of AI and big data for asset pricing.

---

[4] Alternative data are gathered from non-traditional sources (i.e. individuals, businesses and sensors) and come in the form of (un)structured data (e.g., data from social media, web searches, transaction data, corporate data, satellites, geolocation etc.).

5.3 Market Microstructure

The rapid growth of decentralised markets (DEX) has created a unique market structure where Automated Market Makers (AMMs) coexist with traditional centralised exchanges (CEX) and are theoretically connected through arbitrage mechanisms. However, the dynamics of arbitrage in DEX differ significantly due to the AMMs (Foley et al., 2022; Park, 2023).[5] This structural difference can lead to price lags in AMMs, particularly during periods of high volatility, creating unique arbitrage opportunities that traders can exploit between CEX and DEX. Despite arbitrage being key to market efficiency, there are several limiting factors that prevent full informational efficiency as has been shown in traditional markets (Brav et al., 2010). The gap in literature is to identify limitations of arbitrage in DEX, its materiality and constraints on market efficiency by capitalising on AI and big data. The implication of this would further inform the need for continuous financial adaptation.

Price discovery process could be enhanced via the mechanism of information flow, which has received little attention in this era of AI and big data. Theoretically, all available information should be incorporated into asset prices. Thus, future research should capitalise on the benefits of using AI and alternative conditioning data for a comprehensive information dissemination. This would contribute to a timely and accurate price discovery process and an efficiently distributed market.

The dynamics and evolution in financial markets mean there is a high possibility of adverse price movements. Market risk is more prominent and inherent to other risk profiles; such that whenever there are fluctuations in asset prices, the whole financial market gyrates. There is a potential to explore advanced models for predicting market risk by capitalising on big data and AI which is currently underexplored. This is however relevant as the implications of using AI and big data for assessing market risk is a stable financial market.

5.4 Theoretical Advancement

Classical asset pricing models have been criticised continually because the theories and assumptions do not align with the practicality of financial markets. Since early 1900s, there have been empirical advancements aimed at bridging these criticisms. With an overwhelming shift to AI-based models and big data in asset pricing, there is a need for theoretical advancement as AI models are "neither an empirical panacea nor a substitute for economic theory" (Bagnara, 2024). Though AI introduces new paradigms that classical asset pricing theories do not accommodate, there is no theoretical intuition. Altogether, the implication of a

[5] Prices in AMM are set as a ratio of tokenised assets in the liquidity pool.

theoretically advanced AI-based asset pricing model(s) includes improved market efficiency hypotheses and an alignment to financial principles.

6. Conclusion

Motivated by the current wave of AI and big data, we adopt a systematic literature review approach to collate papers from WoS. We systematically review and comprehensively analyse 81 papers on AI and asset pricing, 53 papers on big data and asset pricing, and 24 papers on big data, AI and asset pricing, using bibliometric and thematic analyses. This paper improves on existing reviews which are more narrowly focused on AI and asset pricing and dominantly adopt bibliometric analysis.

There is a surge in the literature in the past 5 years, geographically centred on data from China and US with extensive advancements in factor models and AI-based asset pricing models. The synthesised literature communicates the contribution of AI and big data to enhancing factor analysis, accurate forecasting/prediction models and effective accommodation of market microstructure dynamics leading to optimised portfolio performance and improved risk management strategies. Inherently, AI-based models are effective in identifying asymmetric characteristics and modelling data complexity by capitalising on the conditioning information from high-frequency and high-dimensional data.

The results also reveal advancements to AI-based models built on the foundational classical asset pricing models. These models are efficient in uncovering data patterns that traditional models miss. For researchers and practitioners, this implies that AI-based asset pricing is becoming a more relevant approach for real time data analysis, accommodating market dynamics. Inherently, AI and big data are challenging traditional asset pricing, reinforcing evolving models and emerging alternative and fluid data. Consequently, for investors, AI-based models are useful for effective risk management and optimising portfolio strategies. Market microstructure can also be better understood through improved price discovery processes and arbitrage modelling, ultimately promoting efficient markets. The findings further imply that asset pricing is adaptive. This suggests that AI and big data are challenging traditional asset pricing, reinforcing evolving models and emerging fluid data. Thus, researchers, policy makers, and practitioners should develop theory-data asset pricing models.

Following the search in WoS, there is the likelihood of new papers with additional insights. Thus, researchers can extend this paper by updating the search period and using papers from multiple databases. AI and big data are not just refining asset pricing but are adapting to the practical nature of data and financial markets. Consequently, future research exploring the use of AI and big data would reshape the concept of asset pricing.

Data Availability Statement

Data is available from Web of Science

References

Ahmed, S., Alshater, M. M., El Ammari, A., & Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, *61*, 101646. https://doi.org/10.1016/j.ribaf.2022.101646

Bartram, S. M., Branke, J., & Motahari, M. (2020). *Artificial intelligence in asset management*. CFA Institute Research Foundation.

Bernstein, P. L. (2012). Capital ideas: the improbable origins of modern Wall Street. John Wiley & Sons.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for information Science and Technology*, *61*(12), 2389-2404. https://doi.org/10.1002/asi.21419

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Brav, A., Heaton, J. B., & Li, S. (2010). The limits of the limits of arbitrage. *Review of Finance*, *14*(1), 157-187. https://doi.org/10.1093/rof/rfp018

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

Daugaard, D. (2020). Emerging new themes in environmental, social and governance investing: a systematic literature review. *Accounting & Finance*, *60*(2), 1501-1530. https://doi.org/10.1111/acfi.12479

Daugaard, D., Jia, J., & Li, Z. (2024). Implementing corporate sustainability information in socially responsible investing: a systematic review of empirical research. *Journal of Accounting Literature*, *46*(2), 238-276. https://doi.org/10.1108/jal-12-2022-0127

Dawadi, S. (2020). Thematic analysis approach: A step by step guide for ELT research practitioners. *Journal of NELTA*, *25*(1-2), 62-71. https://doi.org/10.3126/nelta.v25i1-2.49731

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

Engle, R. F. (2000). The econometrics of ultra-high-frequency data. *Econometrica*, *68*(1), 1-22. https://doi.org/10.1111/1468-0262.00091

Fama, E. F. (1965). The behavior of stock-market prices. *The journal of Business*, *38*(1), 34-105. https://doi.org/10.1086/294743

Fama, E. F. (1970). Efficient capital markets. *Journal of finance*, *25*(2), 383-417. https://doi.org/10.2307/2325486

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *Journal of finance*, *47*(2), 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

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. https://doi.org/10.1016/0304-405x(93)90023-5

Fama, E. F., & French, K. R. (1996). Multifactor explanations of asset pricing anomalies. *The Journal of Finance*, *51*(1), 55-84. https://doi.org/10.1111/j.1540-6261.1996.tb05202.x

Fama, E. F., & French, K. R. (2004). The capital asset pricing model: Theory and evidence. *Journal of Economic Perspectives*, *18*(3), 25-46. https://doi.org/10.1257/0895330042162430

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of financial economics*, *116*(1), 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

Foley, S., O'Neill, P., & Putnins, T. (2022). *Can markets be fully automated? evidence from an "automated market maker"*. https://doi.org/10.2139/ssrn.4459924

Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big data in finance. *The Review of Financial Studies*, *34*(7), 3213-3225. https://doi.org/10.1093/rfs/hhab038

Goodell, J. W., Kumar, S., Lim, W. M., & Pattnaik, D. (2021). Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis. *Journal of Behavioral and Experimental Finance*, *32*, 100577. https://doi.org/10.1016/j.jbef.2021.100577

Goodhart, C. A., & O'Hara, M. (1997). High frequency data in financial markets: Issues and applications. *Journal of Empirical Finance*, *4*(2-3), 73-114. https://doi.org/10.1016/s0927-5398(97)00003-0

Hasbrouck, J. (2004). Empirical market microstructure. *Economic and Statistical Perspectives on the Dynamics of Trade in Securities Markets*. https://doi.org/10.1093/oso/9780195301649.003.0012

Lee, P. C., & Su, H. N. (2010). Investigating the structure of regional innovation system research through keyword co-occurrence and social network analysis. *Innovation*, *12*(1), 26-40. https://doi.org/10.5172/impp.12.1.26

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Harvard University Press.

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian journal of management*, *45*(2), 175-194. https://doi.org/10.1177/0312896219877678

Maseda, A., Iturralde, T., Cooper, S., & Aparicio, G. (2022). Mapping women's involvement in family firms: A review based on bibliographic coupling analysis. *International Journal of Management Reviews*, *24*(2), 279-305. https://doi.org/10.1111/ijmr.12278

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., & Brennan, S. E. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *British Medical Journal*, *372*. https://doi.org/10.1016/j.jclinepi.2021.02.003

Park, A. (2023). The conceptual flaws of decentralized automated market making. *Management Science*, *69*(11), 6731-6751. https://doi.org/10.1287/mnsc.2021.02802

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). World Scientific. https://doi.org/10.1142/9789814417358_0001

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229. https://doi.org/10.1147/rd.33.0210

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

Tucker, A. (2004). *Our knowledge of the past: A philosophy of historiography*. Cambridge University Press.

Vespignani, J., & Smyth, R. (2024). Artificial intelligence investments reduce risks to critical mineral supply. *Nature Communications*, *15*(1), 7304. https://doi.org/10.1038/s41467-024-51661-7

Walter, C., & Ribière, V. (2013). A citation and co-citation analysis of 10 years of KM theory and practices. *Knowledge Management Research & Practice*, *11*(3), 221-229. https://doi.org/10.1057/kmrp.2013.25

Yang, R., Yu, L., Zhao, Y., Yu, H., Xu, G., Wu, Y., & Liu, Z. (2020). Big data analytics for financial Market volatility forecast based on support vector machine. *International Journal of Information Management*, *50*, 452-462. https://doi.org/10.1016/j.ijinfomgt.2019.05.027

Table 1: Search Strategy

Panel A: Concept and Keywords		
Big Data	*Asset Pric**	*Artificial Intelligence OR AI*
High-frequency data	Efficient market hypothesis	Machine learning
Real time market data	(Capital Asset Pricing Model OR CAPM)	Deep learning
	(Arbitrage Pricing Theory OR APT)	Algorith* trad*
	Factor model*	Robo* analy*

Panel B: Research Questions, Themes and Search String		
Research Question	***Theme***	***Search String***
What is the impact of AI on Asset Pricing?	AI AND Asset Pricing	((SU=("Business" OR "Business Finance" OR "Economics")) AND TS=(("asset pric*" OR "efficient market hypothesis" OR "capital asset pricing model" OR "CAPM" OR "arbitrage pricing theory" OR "APT" OR "factor model*"))) AND TS=(("artificial intelligence" OR "AI" OR "deep learning" OR "machine learning" OR "algorith* trad*" OR "robo* analy*"))
What is the impact of Big Data on Asset Pricing?	Big Data AND Asset Pricing	((SU=("Business" OR "Business Finance" OR "Economics")) AND TS=(("big data" OR "high-frequency data" OR "real time market data"))) AND TS=(("asset pric*" OR "efficient market hypothesis" OR "capital asset pricing model" OR "CAPM" OR "arbitrage pricing theory" OR "APT" OR "factor model*"))
Collectively, how are big data and AI transforming asset pricing?	Big Data AND AI AND Asset Pricing	(((SU=("Business" OR "Business Finance" OR "Economics")) AND TS=(("big data" OR "high-frequency data" OR "real time market data"))) AND TS=(("asset pric*" OR "efficient market hypothesis" OR "capital asset pricing model" OR "CAPM" OR "arbitrage pricing theory" OR "APT" OR "factor model*"))) AND TS=(("artificial intelligence" OR "AI" OR "deep learning" OR "machine learning" OR "algorith* trad*" OR "robo* analy*"))

Note: SU (Research Area); TS (Topic); * is a wildcard for keywords that may have variations (example, model* for modelling, modeling, and models). The search was conducted on 17/09/2024 in WoS. For each research question, a unique search string is used to combine concepts ("AND") and their respective keywords ("OR").

Table 2: Inclusion and Exclusion Criteria

Criteria	**Inclusion**	**Exclusion**
• Literature Type	• Journal Publications	• Retracted Publications • Books • Book Chapters • Dissertations/Theses • Conference proceedings • Editorials
• Language	• English	• Non-English
• Research Area	• Business • Business Finance • Economics	• Health and Medicine • Computer Science • Engineering

Table 3: 10 Most Relevant Author(s) Keywords

AI and Asset Pricing		Big Data and Asset Pricing		Big Data, AI and Asset Pricing	
Author(s) Keywords	**No of Articles**	**Author(s) Keywords**	**No of Articles**	**Author(s) Keywords**	**No of Articles**
Machine Learning	28	Big Data	22	Machine Learning	13
Asset Pricing	14	Machine Learning	15	Big Data	11
Big Data	8	Factor Model*	13	***Neural Network****	***6***
Deep Learning	7	Forecasting	6	Deep Learning	4
Factor Model	6	High-frequency Data	6	Forecasting	4
Neural Networks	***5***	Asset Pricing	5	Artificial Intelligence	3
Stock Returns	5	Deep Learning	4	Asset Pricing	3
Artificial Intelligence	4	Artificial Intelligence	3	Conditional Asset Pricing Model	3
Empirical Asset Pricing	4	Conditional Asset Pricing Model	3	Autoencoder	2
Factor Models	4				

Note: The author(s)' keywords reflect the intended focus of the paper. * is used to truncate keywords: Factor Model* → Factor Model (7) and Factor Models (6); Neural Network* → Neural Network (3) and Neural Networks (3). The keywords were updated for AI with "*neural network**".

Table 4: Bibliometric Analysis and Unit of Analysis

Bibliometric Analysis	Unit of Analysis
Bibliographic Coupling	Documents (Minimum of 10 shared references)
Co-citation Analysis	Cited References (Minimum of 10 references cited together)
Historiograph	Documents (presented as 1st Author and Year of publication)
Keyword Co-occurrence	Author(s)' Keywords

Table 5: Descriptives of the Papers used in the Systematic Literature Review

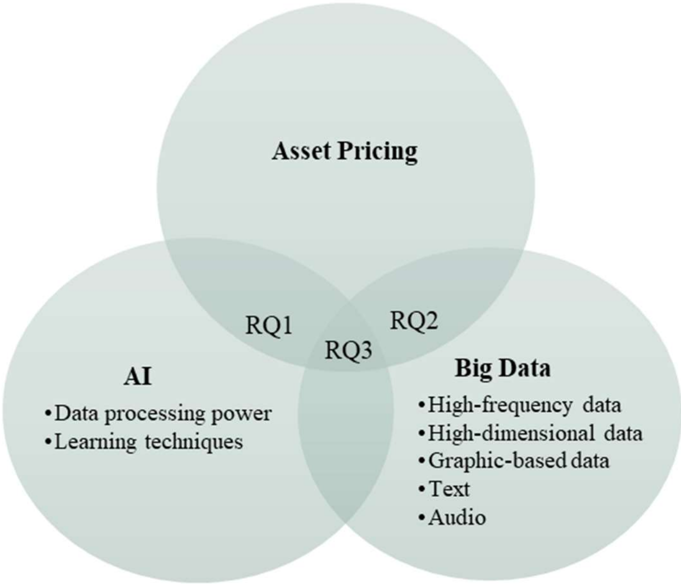
	AI and Asset Pricing	Big Data and Asset Pricing	Big Data, AI and Asset pricing
Panel A: Main Information			
Time span	1994:2024	2006:2024	2006:2024
Annual Growth rate of publications	9.45%	12.25%	6.29%
Average citations per paper (per year)	18.49 (2.621)	20.75 (3.455)	31.04 (5.582)
Document Types			
Article	79	53	24
Review	2	-	-
Author(s) Collaboration			
Single-authored papers	13	5	2
Multiple-authored papers	68	48	22
Database			
Web of Science	81	53	23
Other (SciELO Citation Index)	-	-	1
Total Papers	***81***	***53***	***24***

Panel B: 10 Most Cited Papers						
	Author(s) (Year)	**Total Citations**	**Author(s) (Year)**	**Total Citations**	**Author(s) (Year)**	**Total Citations**
	Heaton et al. (2017)	320	Heaton et al. (2017)	320	Heaton et al. (2017)	320
	Refenes et al. (1994)	138	Gu et al. (2021)	123	Gu et al. (2021)	123
	Gu et al. (2021)	123	Shang et al. (2019)	100	L. Y. Chen. et al. (2024)	58
	Cao et al. (2006)	113	Fan et al. (2016)	71	Kim and Swanson (2018)	59
	Ho et al. (2011)	110	L. Y. Chen et al. (2024)	58	Coulombe et al. (2022)	38
	Weller (2018)	105	Lee and Mykland (2012)	56	Karolyi and Van Nieuwerburgh (2020)	31
	L. Y. Chen et al. (2024)	58	Masoliver et al. (2006)	51	Costa et al. (2021)	25
	Fan et al. (2017)	44	Coulombe et al. (2022)	38	Weigand (2019)	24
	Bai and Ng (2019)	37	Dai et al. (2019)	28	Shephard and Xiu (2017)	17
	Giglio et al. (2021)	34	Costa et al. (2021)	25	Centanni and Minozzo (2006)	14
			Kim and Fan (2019)	25		

Panel C: Annual Publications	1994	1995	1996-2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
AI and Asset Pricing	1	1	0	1	1	0	0	1	0	0	2	0	1	0	1	1	4	1	3	8	13	11	21	10
Big Data and Asset Pricing	0	0	0	0	0	1	0	0	0	1	0	1	1	1	0	1	2	1	6	7	5	3	18	5
Big Data, AI and Asset pricing	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	1	1	2	2	4	8	3

Note: *The search is conducted in "All Databases" in WoS, capturing a paper from SciELO Citation Index. Total Citations is the number of times the paper has been cited in the WoS database.*

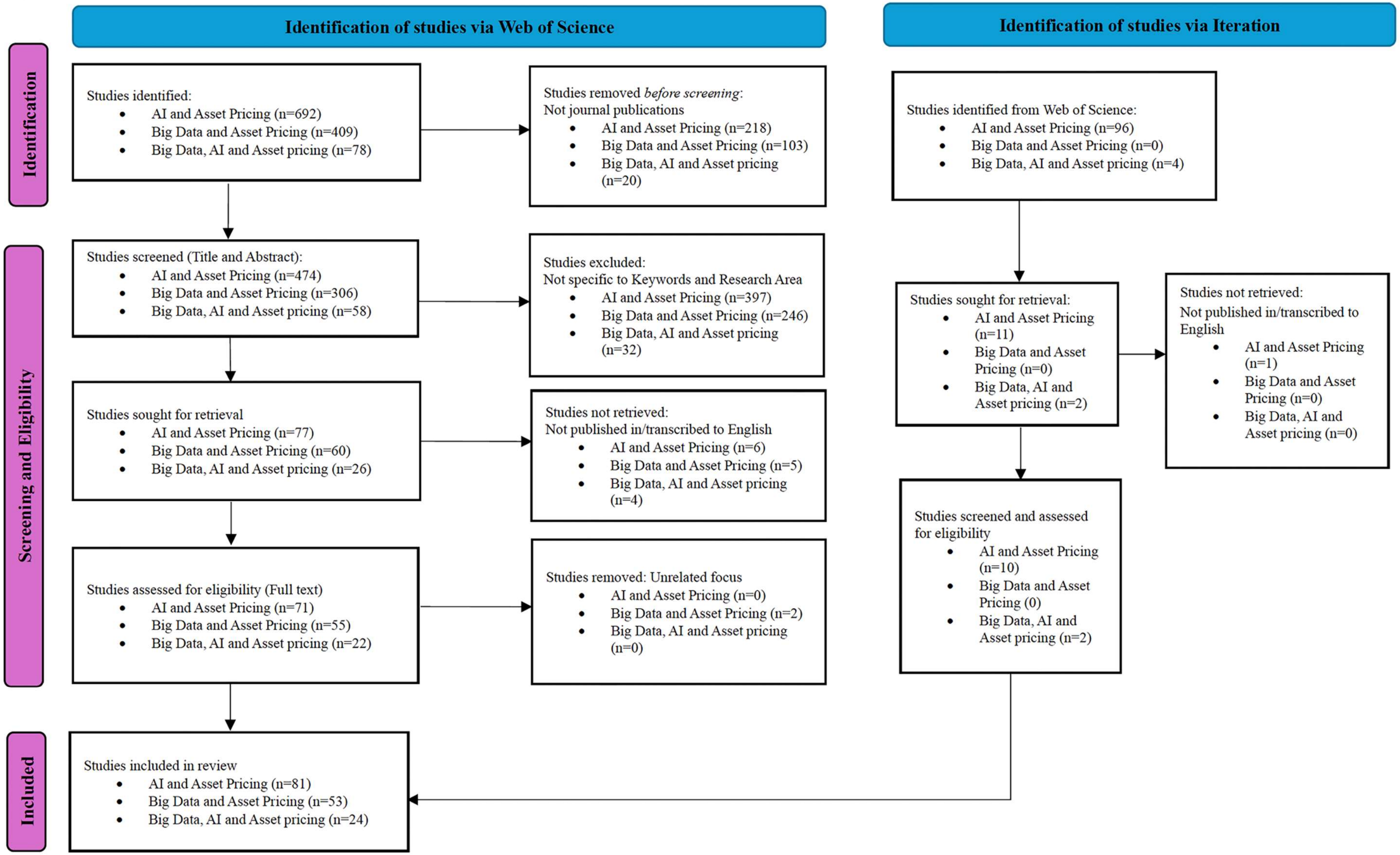
Figure 1: Thematic intersections of Big Data, AI, and Asset Pricing



RQ1: What is the impact of AI on Asset Pricing?
RQ2: What is the impact of Big Data on Asset Pricing?
RQ3: Collectively, how are big data and AI transforming asset pricing?

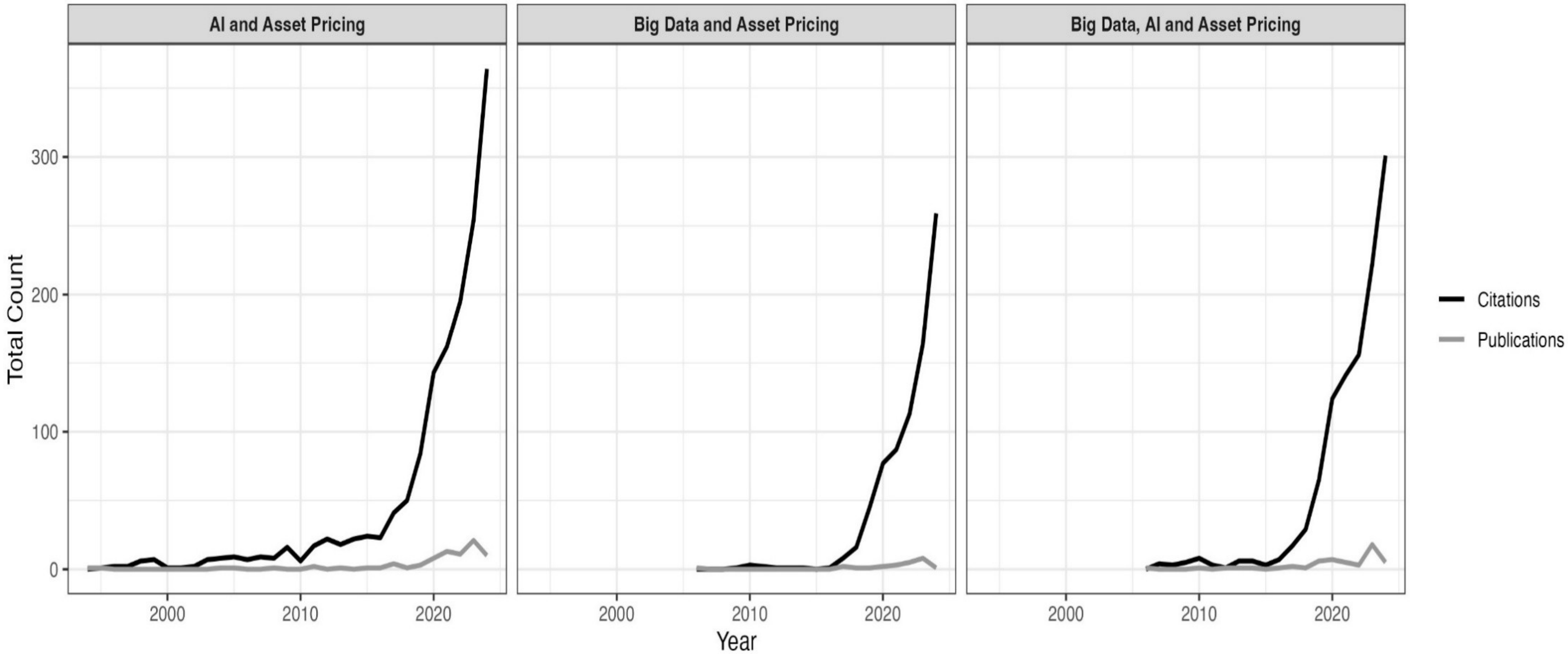
Note: RQ (Research Question). Big data is an expanded information set, offering alternative predicting variables for asset pricing. AI has data processing functions and learning techniques to handle and draw insights from varying, complex and large datasets. Their intersection highlights a transformative impact on asset pricing.

Figure 2: PRISMA Flow Diagram of Included Papers



Note: The papers were identified in WoS on 17/09/2024.

Figure 3: Annual Total Publication and Citations



Note: The total count for citations and publications over time is captured in WoS as Citation Report

Figure 4: Bibliographic Coupling of AI and Asset Pricing

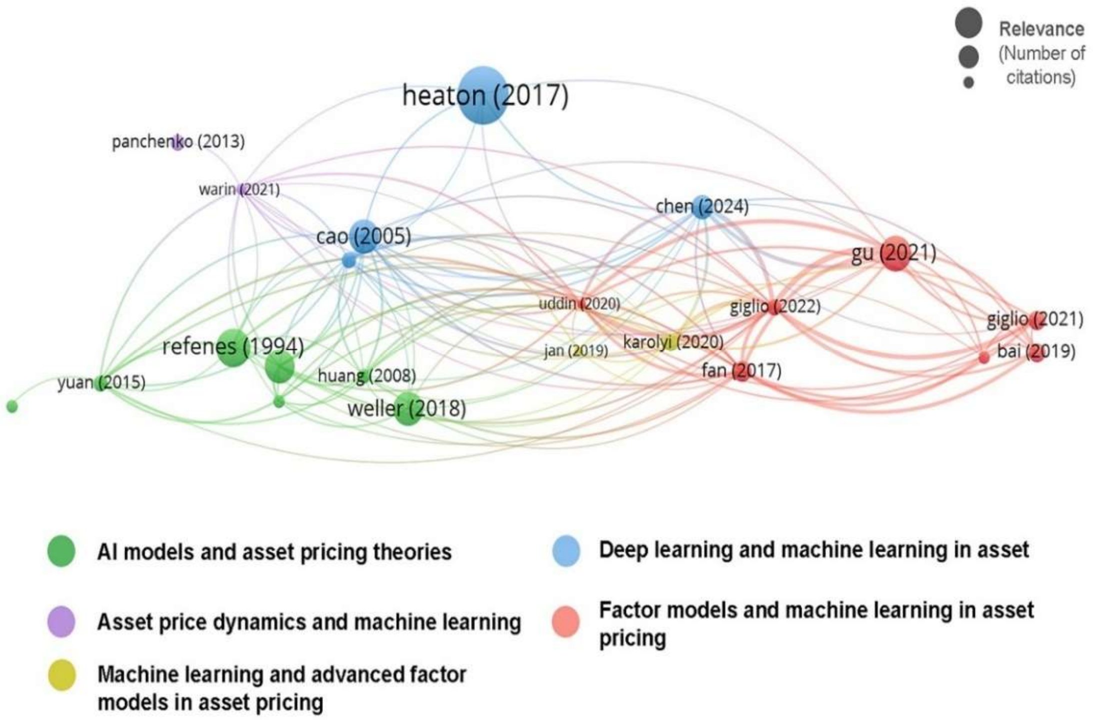
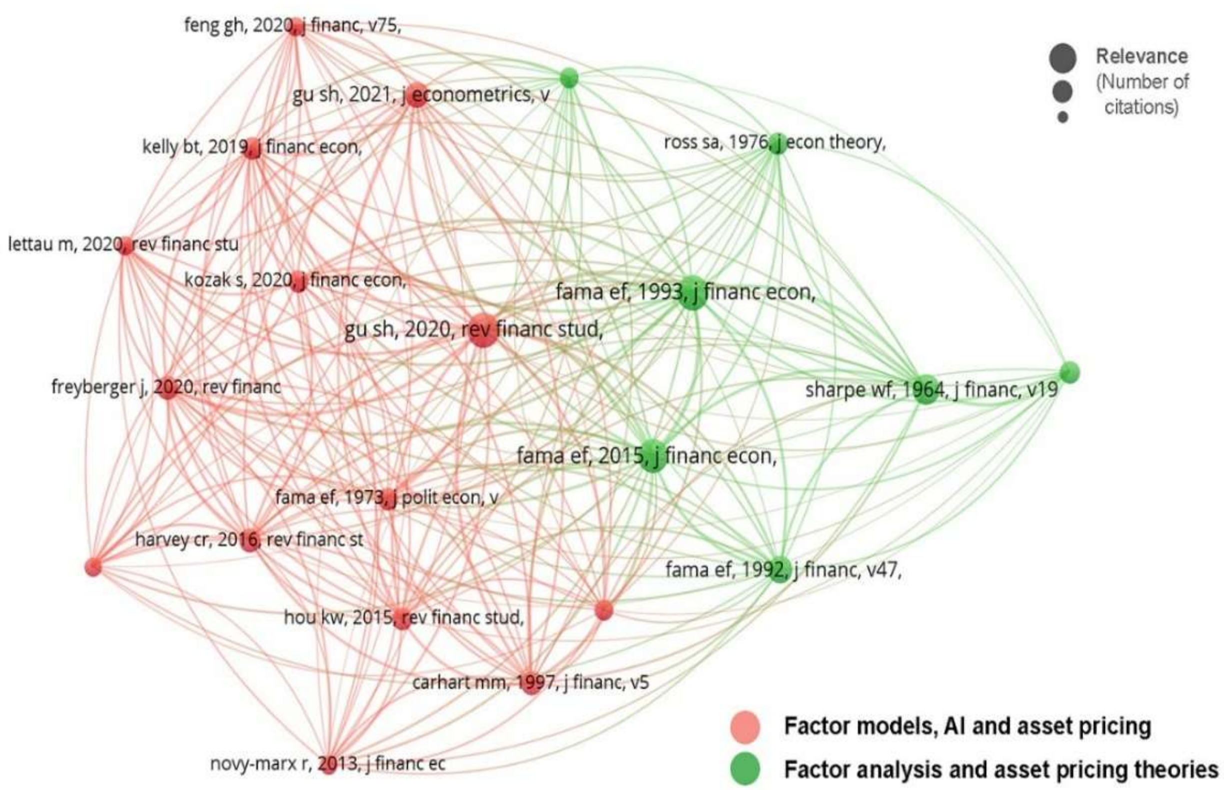
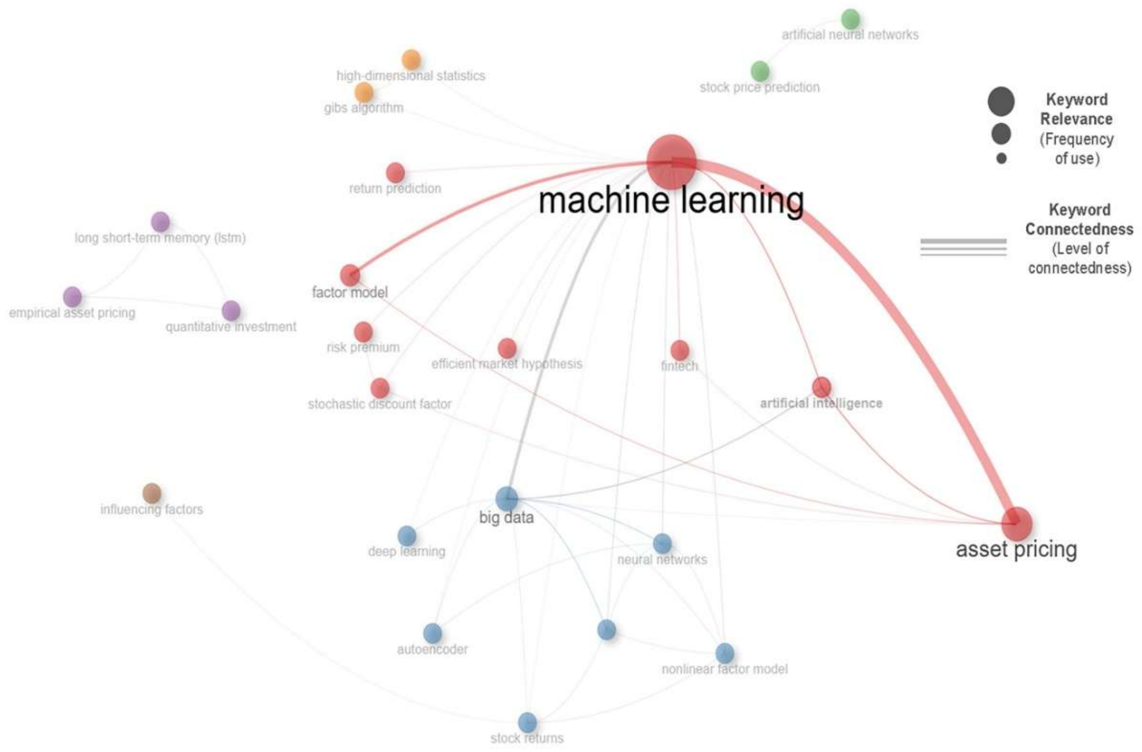


Figure 5: Co-citation analysis of AI and Asset Pricing



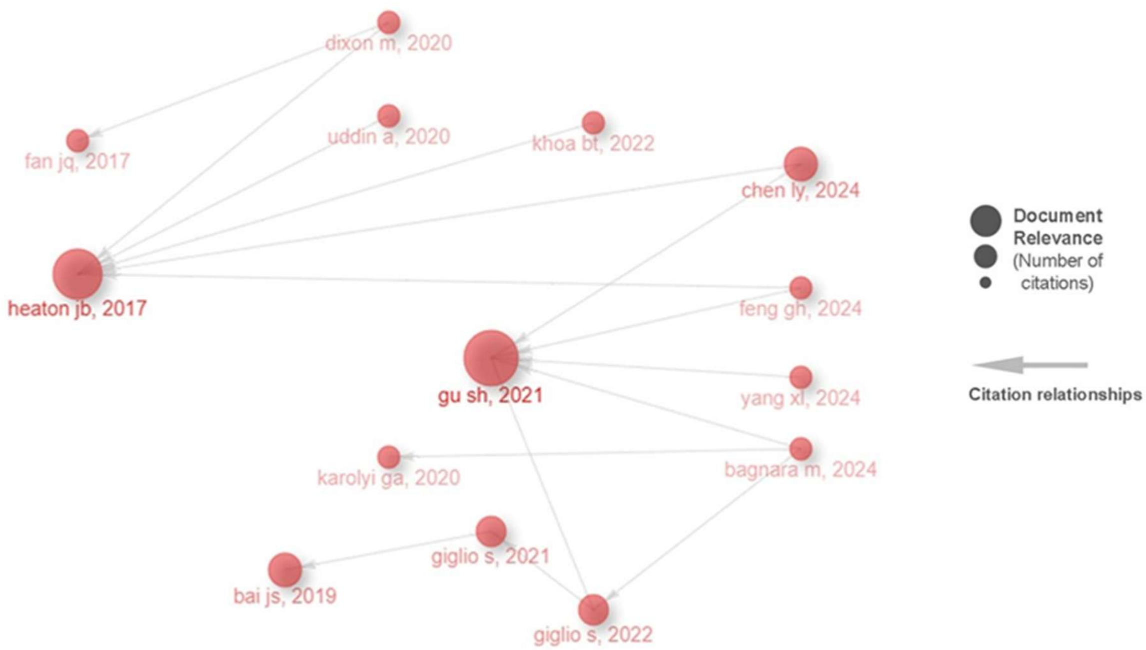
Note: Using a minimum of 10 citations of a cited reference, only 21 cited references meet this threshold out of 2825 cited references.

Figure 6: 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 7: Historiograph of AI and Asset Pricing



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

Figure 8: Bibliographic Coupling of Big Data and Asset Pricing

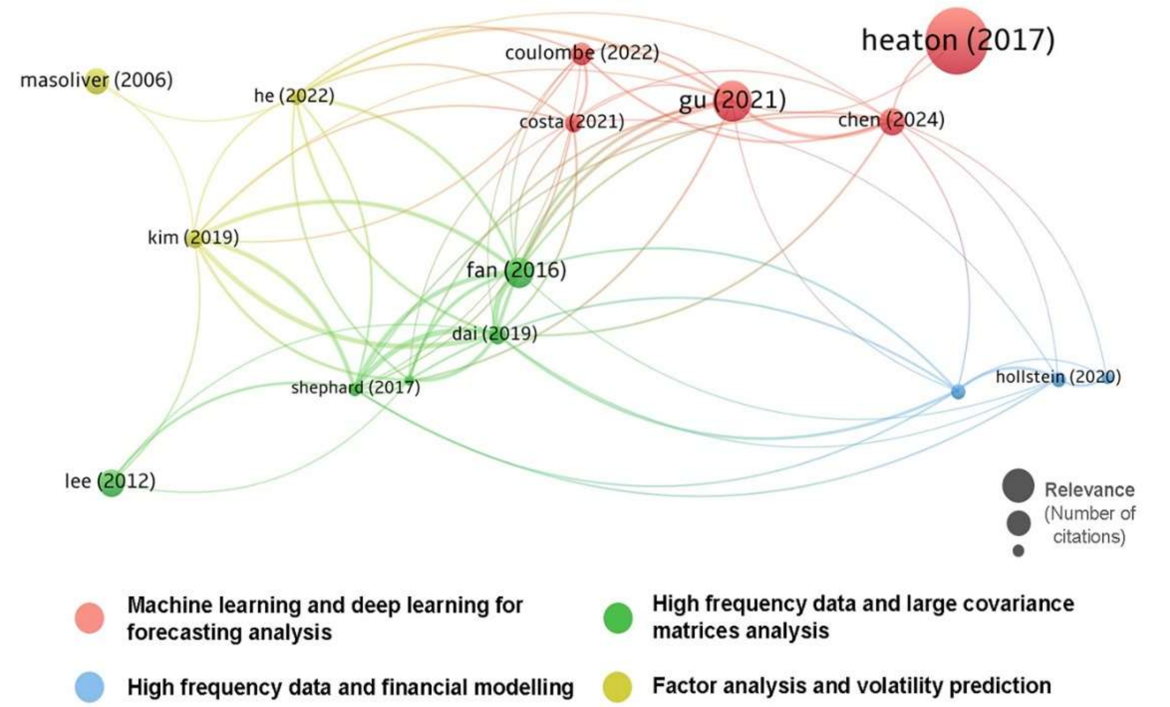
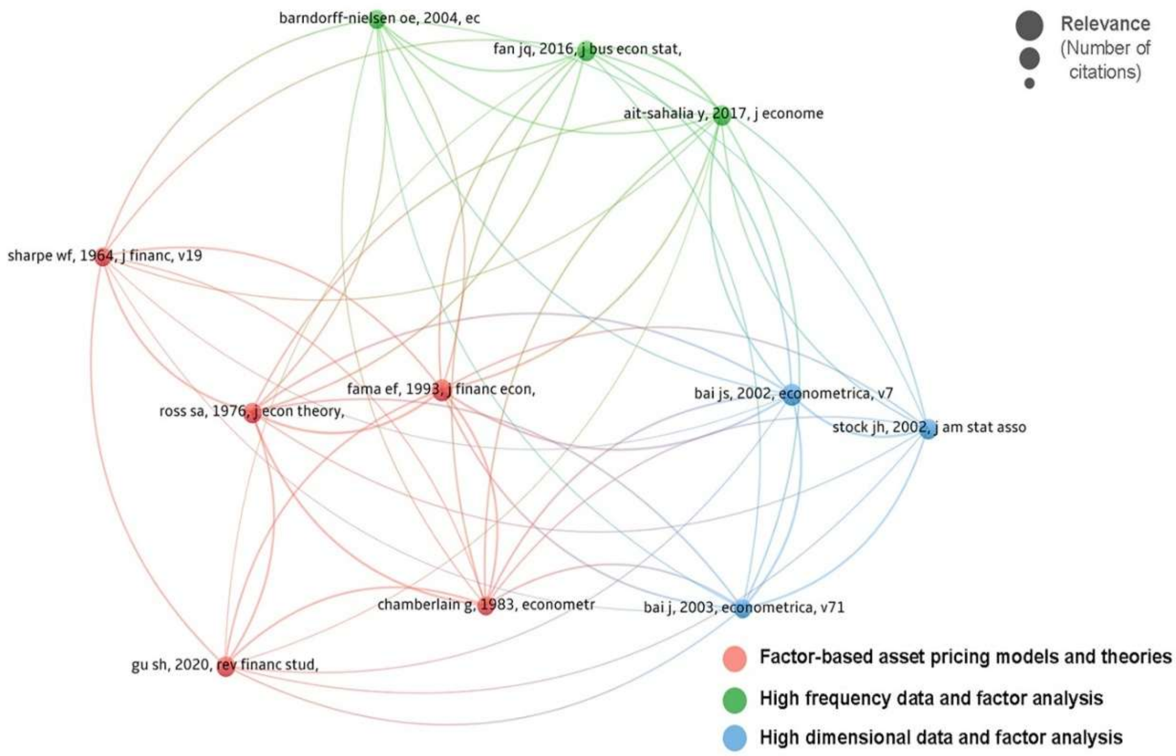
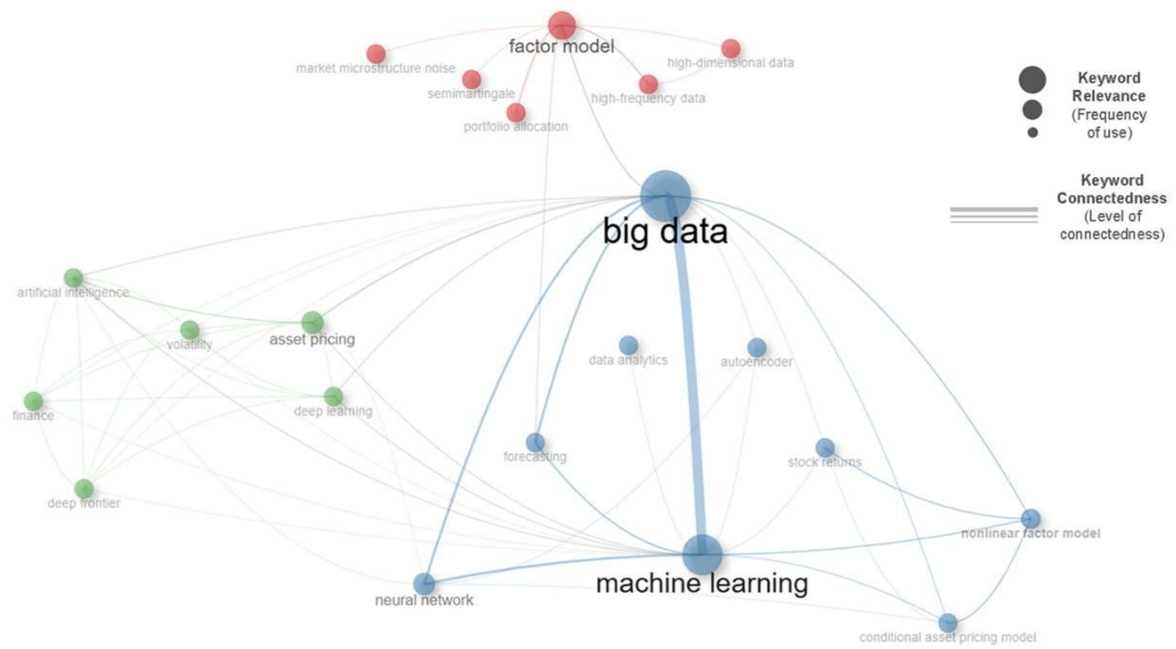


Figure 9: Co-citation analysis of Big Data and Asset Pricing



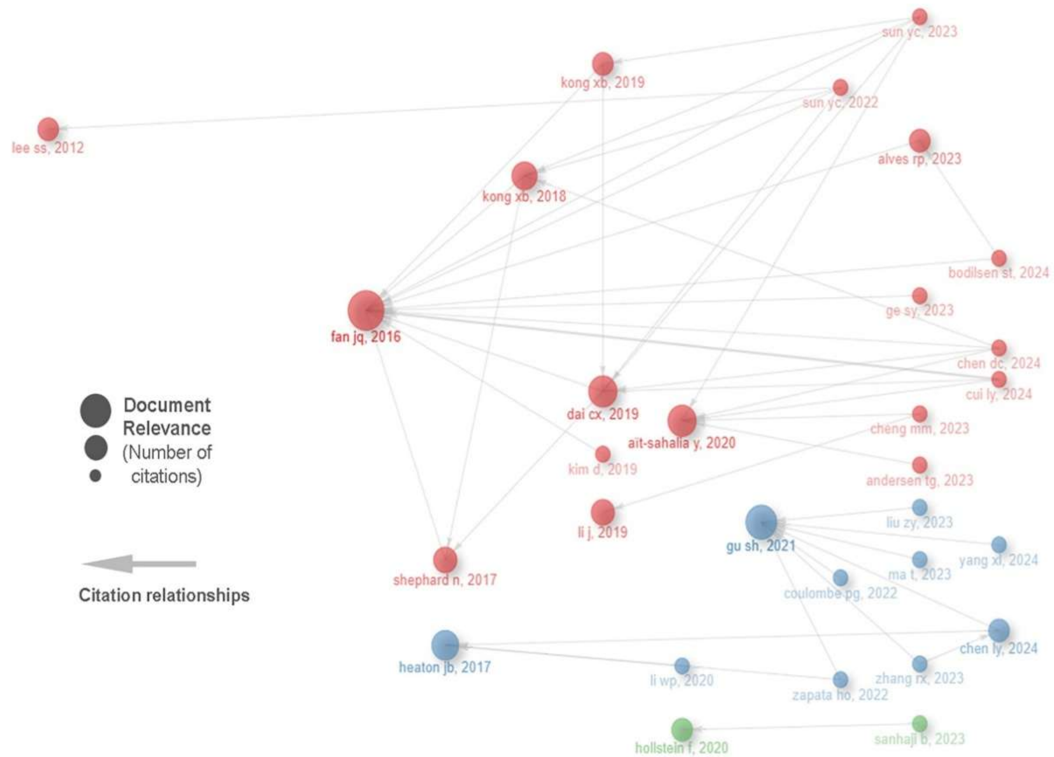
Note: Using a minimum of 10 citations of a cited reference, only 18 cited references meet this threshold out of 1899 cited references.

Figure 10: 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 11: Historiograph of Big Data and Asset Pricing



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

Figure 12: Bibliographic Coupling of Big Data, AI and Asset pricing

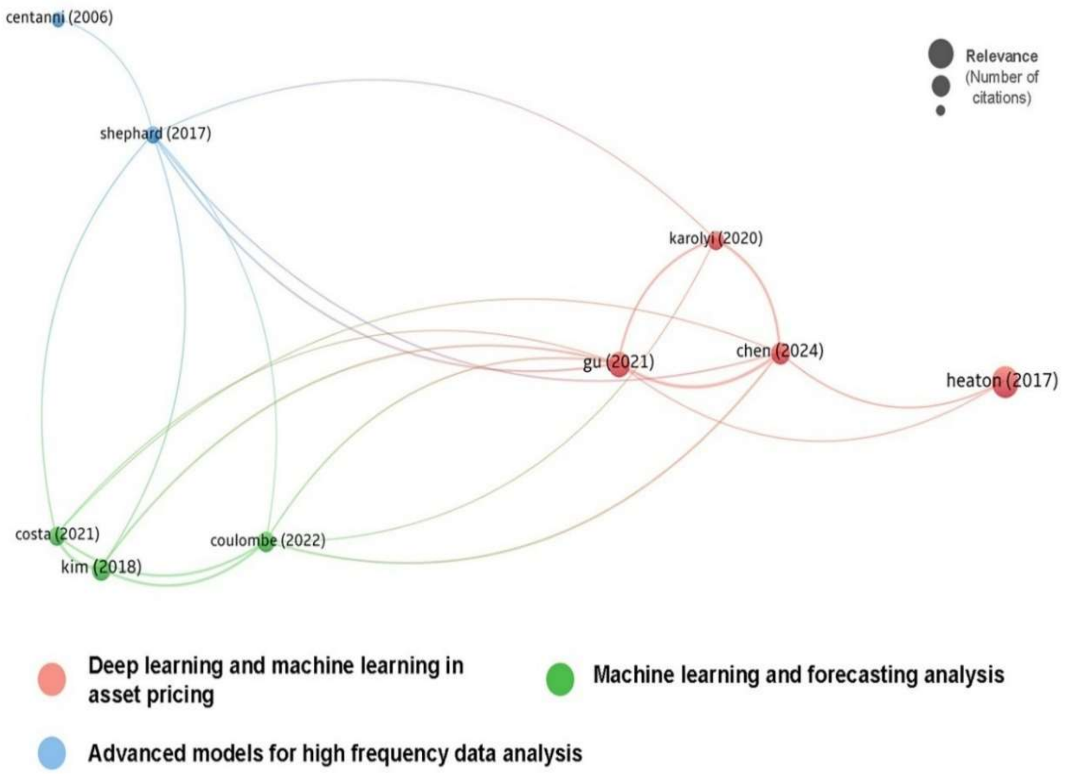
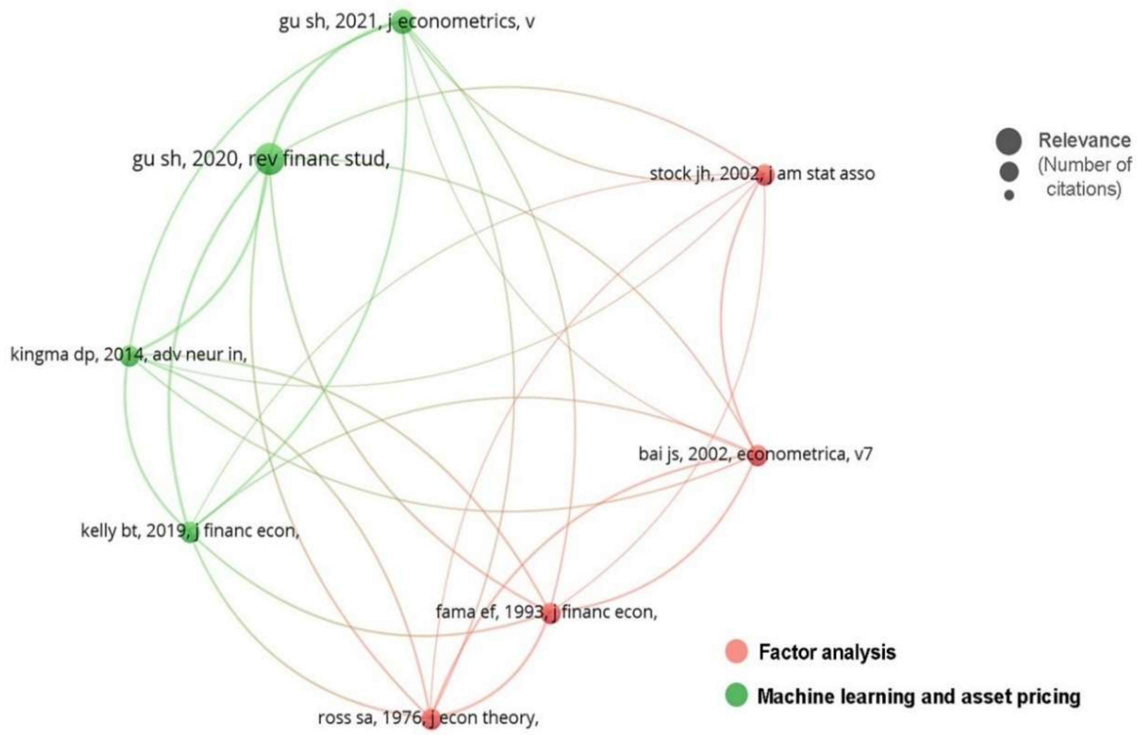
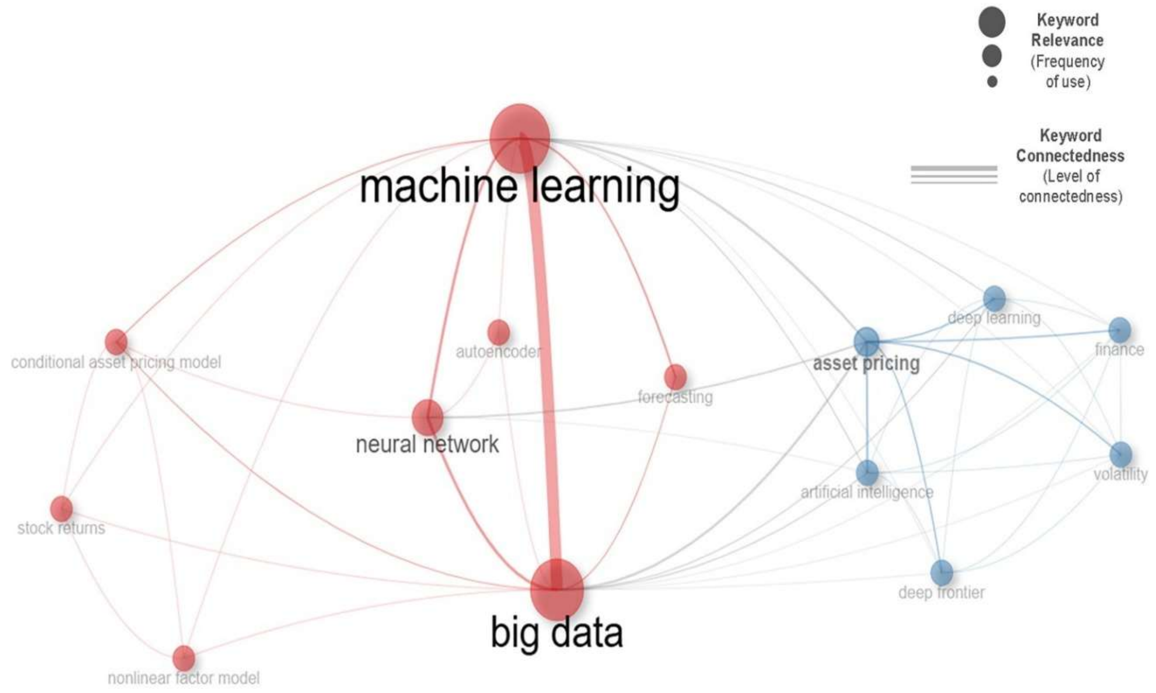


Figure 13: Co-citation analysis of Big Data, AI and Asset pricing



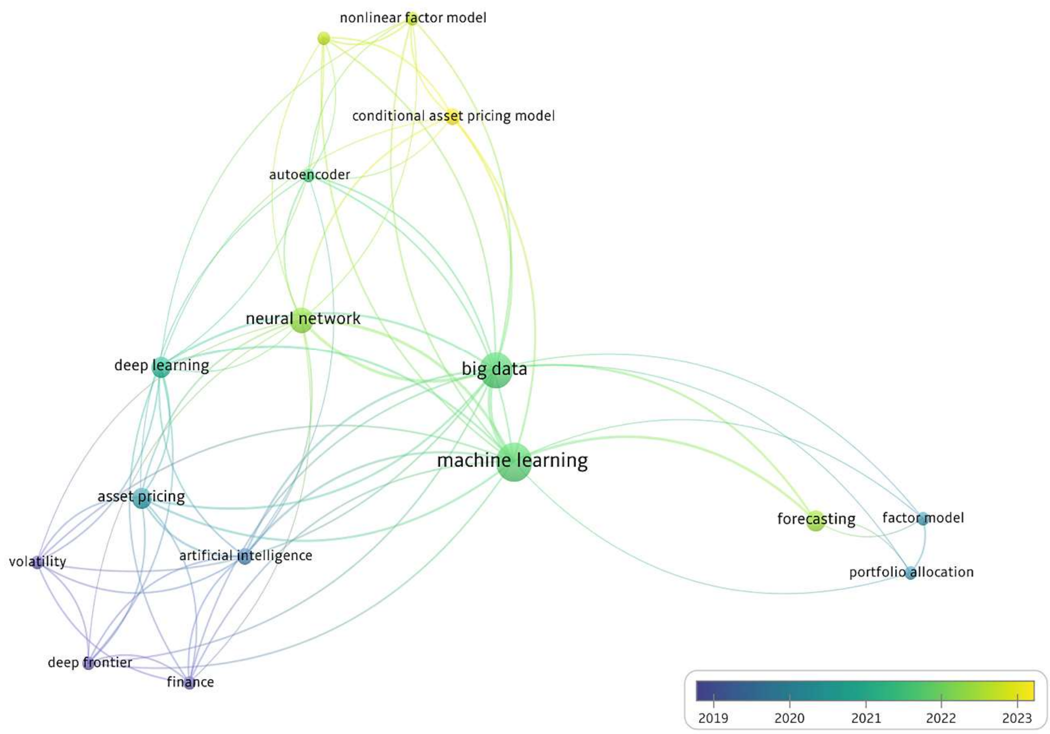
Note: Using a minimum of 5 citations of a cited reference, only 8 cited references meet this threshold out of 1964 cited references.

Figure 14: Keyword Co-occurrence of Big Data, AI and Asset pricing



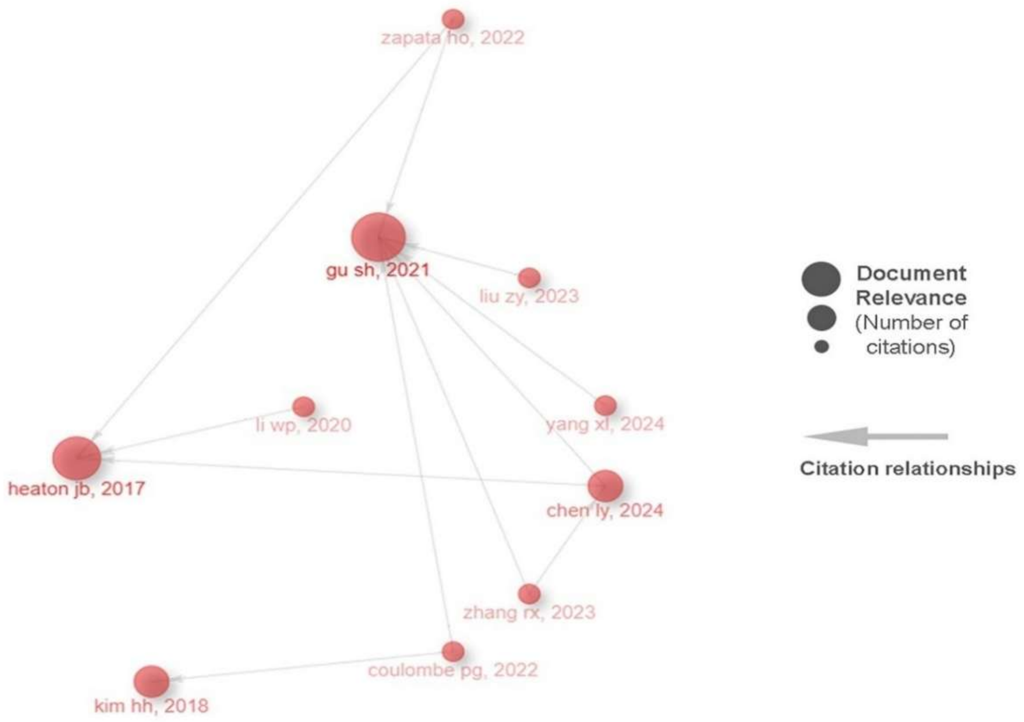
Note: The figure represents how often author(s) keywords appear together in the literature. The circles are keywords, edges show the level of keywords connectedness, and the colour codes are clusters, representing a theme. The thickness of the edges show the strength in connectedness and the size of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

Figure 15: Timeline of Keywords in Big Data, AI and Asset pricing



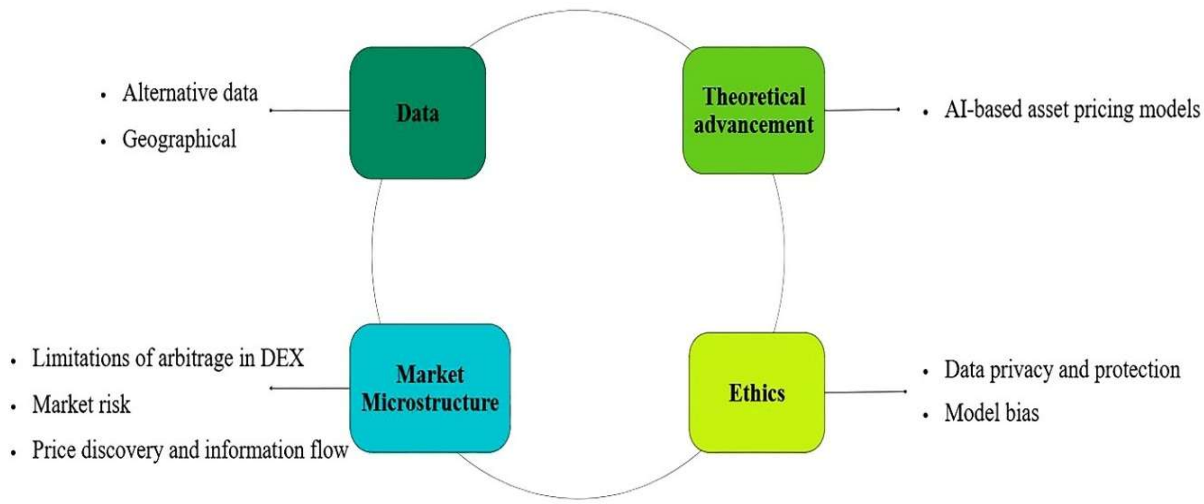
Note: Using a minimum of 2 occurrences for a keyword, there are 17 author(s) keywords that meet the threshold out of 103 author(s) keywords.

Figure 16: Historiograph of Big Data, AI and Asset pricing



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

Figure 17: Gaps in the literature and directions for future studies



Appendix A: Thematic Analysis

Table A1: Identification of Themes

Theme	Sub-themes	Codes	Author(s) (Year)
		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) • Fang et al. (2023) • Rui (2023) • Han (2021) • Cui et al. (2024)
		AI models are effective in enhancing factor analysis	• 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)
	Improve prediction/forecasting accuracy	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)
		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)
		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)
Improve Predictive/ Forecasting Accuracy	Capture complex and nonlinear relationships	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)
		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)
	Robust in handling data	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)
		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)
		AI models are effective in the long run	• Maehashi and Shintani (2020) • French (2017) • Azevedo et al. (2023)

Model Robustness and Adaptability	Adaptability to market conditions	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)
		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)
	Application in asset pricing theories and anomalies	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)
		AI models are effective in exploring asset pricing or market anomalies	• Caporale et al. (2016) • Panchenko et al. (2013)
Enhance financial and investment decision making	Improve risk management and portfolio optimisation	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)
		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) • 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			
		AI/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)

		AI/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)
	Outperforming AI/Hybrid models	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)
	Model nonlinear and complex relationships	AI/hybrid models are effective in handling large, complex, and incomplete data	• Camacho and Lopez-Buenache (2023) • Dai et al. (2019)
		Nonlinear models significantly improve forecasting accuracy	• Coulombe et al. (2022) • Pan et al. (2023)
	Robustness and Reliability	AI/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		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)
	Leveraging data for improved accuracy	The marginal effect of big data is positive and significant, improving forecasting accuracy	• Coulombe et al. (2022)

		Enhanced factor models optimise portfolio performance and allocation through advanced factor selection and estimation	• Sun et al. (2023) • Fan et al. (2016)
		AI/hybrid and 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)
		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 forecasting accuracy	• Bae (2024)
Optimising Factor Analysis	Factor selection and asset pricing	Imposing a factor structure stabilises risk and uncertainty	• Oomen (2010)
	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 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)
Portfolio Optimisation and Risk Management		Hybrid models effectively reduce volatility and lower out-of-sample risks	• Kim and Fan (2019) • Alves et al. (2023)
	Advanced risk assessment	Machine learning improves risk premium measurement	• Weigand (2019)
		Short training models have higher success rates	• Martínez et al. (2020)
	Training data and sample size	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)
	Optimising arbitrage and market alignment	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)
		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)
		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)
Market Dynamics and Macroeconomic Influences	Idiosyncratic factors	Idiosyncratic betas are time-varying	• Cheng et al. (2023)
Big Data, AI and Asset pricing			
		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)
	Outperforming AI/Hybrid models	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)
		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)
		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)

	Model nonlinear and complex relationships	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)
Improve Prediction/ Forecasting accuracy	Model robustness and reliability	AI models are robust and reliable for forecasting and prediction	• L. Y. Chen et al. (2024) • Liu et al. (2023)
		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 augmented 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.

Supplementary Appendix

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

Supplementary Appendix A: Methodological Review and Data Scope of Included Papers

Table A1: Methodological Review of AI in Asset Pricing

Reference	AI Model	Theoretical and Empirical Framework	Functionality of AI Model	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. 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. (2004)	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. (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 A2: 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., CRSP, NYSE, NASDAQ, AMEX, CRSP, S&P100, S&P500)	11	9	5
	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
	Theoretical model	2	1	1
		81	**53**	**24**

Note: This table presents the data scope in the respective data corpus. Papers that use data across countries are captured under "global data". Papers that developed theoretical models and used synthetic data or were silent about the data are captured under "theoretical model".

Supplementary Appendix B: Bibliometric Analysis

Figure B1: Bibliographic Graphs for AI and Asset Pricing

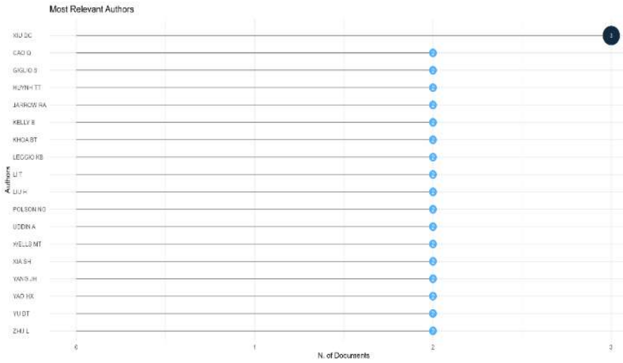


Figure S.1.1: Most Relevant Authors

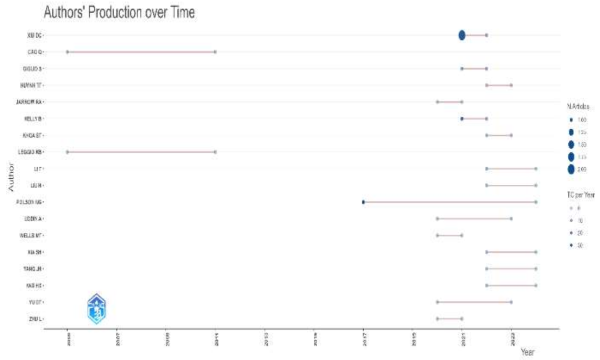


Figure S.1.2: Most Relevant Authors' Production over time

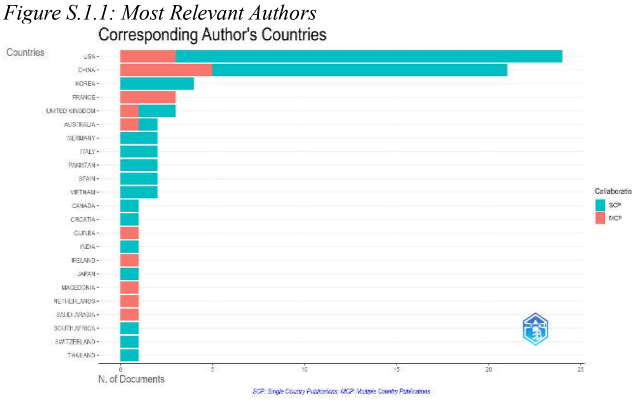


Figure S.1.3: Corresponding Author's Countries

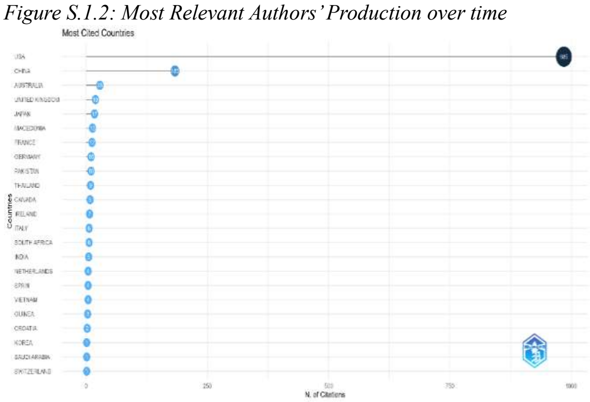


Figure S.1.4: Most Cited Countries

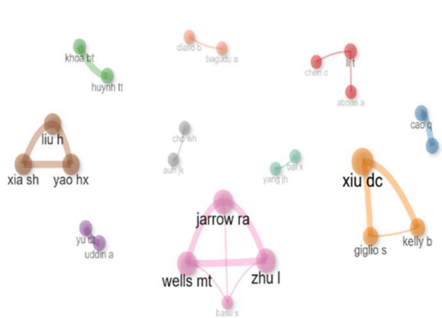


Figure S.1.5: Author Collaboration

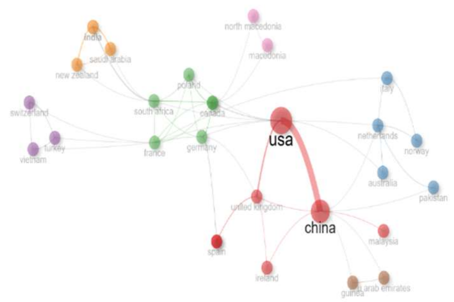


Figure S.1.6: Country Collaboration

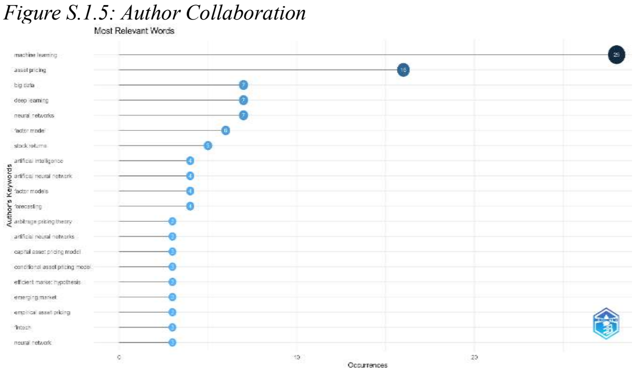


Figure S.1.7: Most Frequent Words

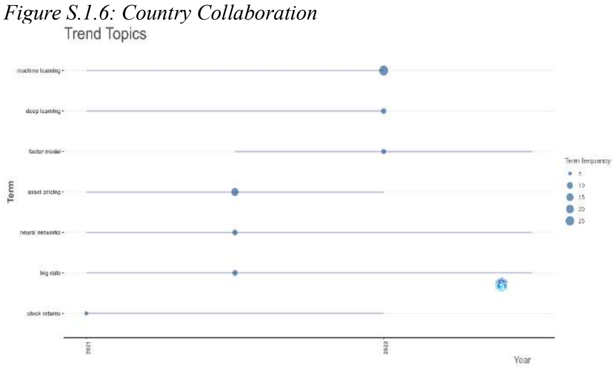


Figure S.1.8: Trend Topics

Note: Most Relevant Authors presents authors with at least 2 publications (18 authors). Authors' Production over time presents the most relevant authors and their production over time. Corresponding authors countries presents 23 corresponding authors. Most Cited countries represent most cited corresponding author-affiliated countries. Most Relevant Words and Trend Topics were analysed using Authors' Keywords. Entries of Switzerland may differ because on the day of downloading the data for the analysis (15/01/2025), Weigand (2019) was not in Web of Science (though the customer service had proved it was in their database); it had to be done manually for other analysis (see Supplementary Table 5) except for the graphs. That is the limitation of the software (VosViewer and Biblioshiny); it does not analyse merged files from different databases.

Figure B2: Bibliographic Graphs for Big Data and Asset Pricing

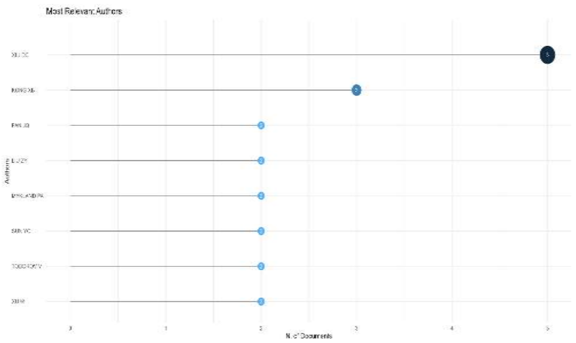


Figure S.2.1: Most Relevant Authors

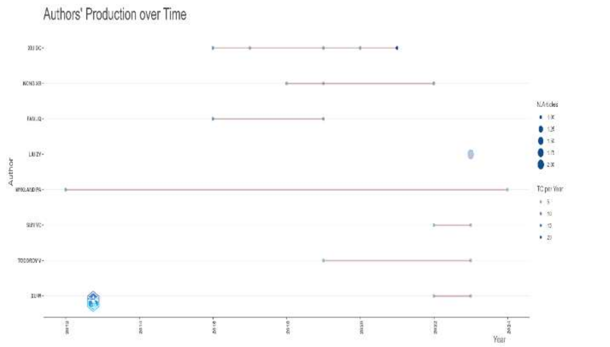


Figure S.2.2: Most Relevant Authors' Production over time

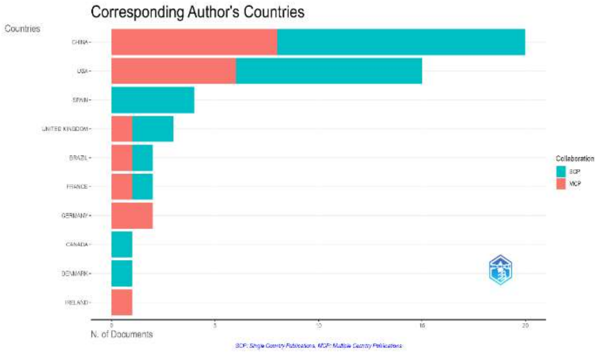


Figure S.2.3: Corresponding Author's Countries

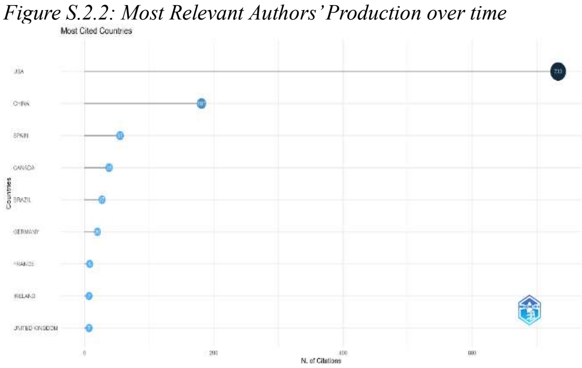


Figure S.2.4: Most Cited Countries

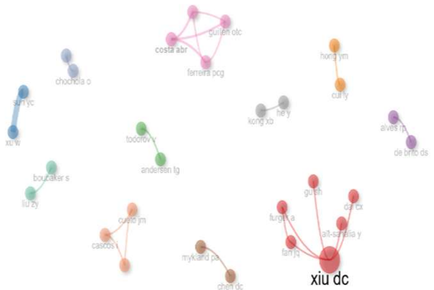


Figure S.2.5: Author Collaboration

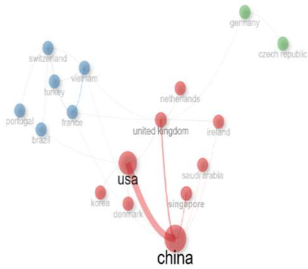


Figure S.2.6: Country Collaboration

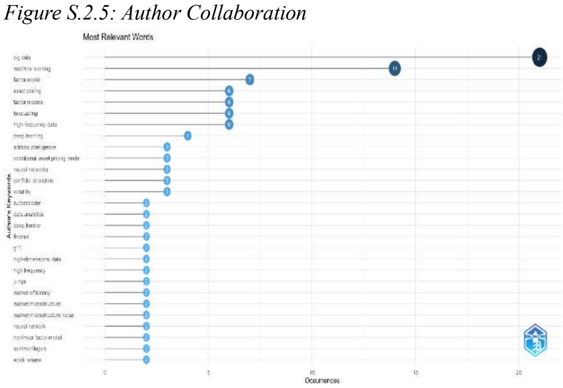


Figure S.2.7: Most Frequent Words

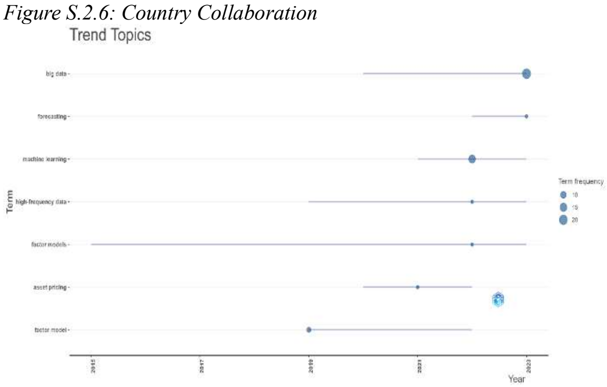


Figure S.2.8: Trend Topics

Note: The Most Relevant Authors represents 8 authors with more than 2 papers. Authors' Production over time is the most relevant authors and their production over time. There are 10 corresponding authors' countries. Most Cited countries represent the corresponding author-affiliated countries (Denmark has 0 citations but there is a publication from Switzerland that has 24 publications which is presented in Supplementary Table 5; (Weigand, 2019)). Most Relevant Words and Trend Topics graphs were analysed using Authors' Keywords. Weigand (2019) was not in Web of Science (though the customer service had proved it was in their database); it had to be done manually for other analysis (see Supplementary Table 5) except for the graphs. That is the limitation of the software (VosViewer and Biblioshiny); it does not analyse merged files from different databases.

Figure B3: Bibliographic graphs for Big Data, AI and Asset Pricing

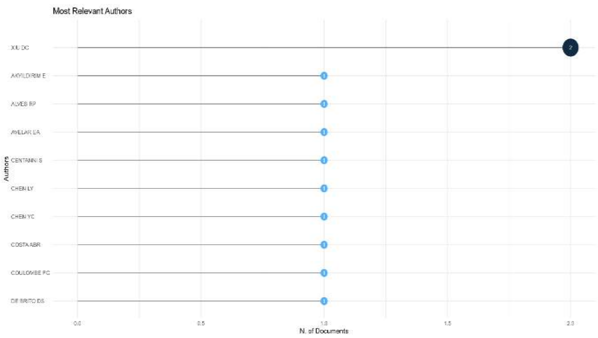


Figure S.3.1: Most Relevant Authors

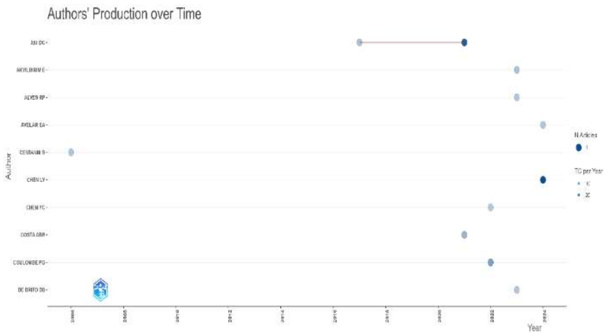


Figure S.3.2: Most Relevant Authors' Production over time

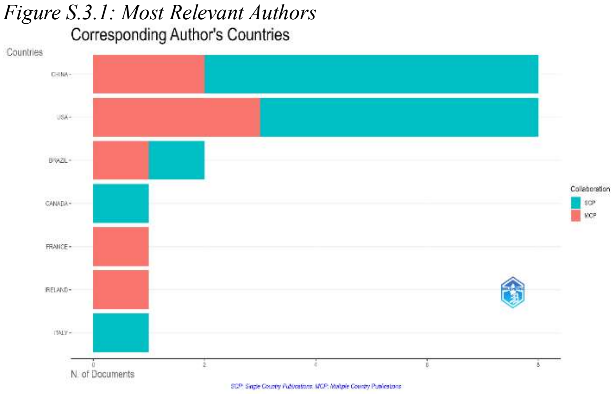


Figure S.3.3: Corresponding Author's Countries

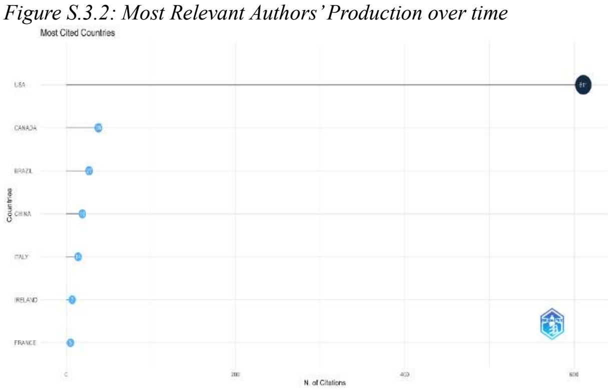


Figure S.3.4: Most Cited Countries

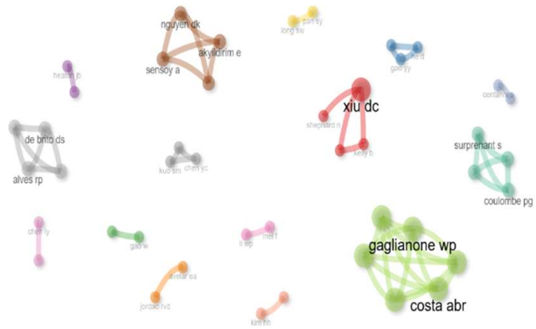


Figure S.3.5: Author Collaboration

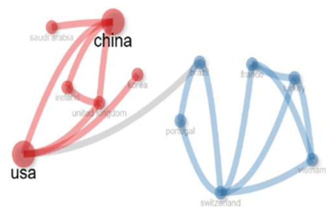


Figure S.3.6: Country Collaboration

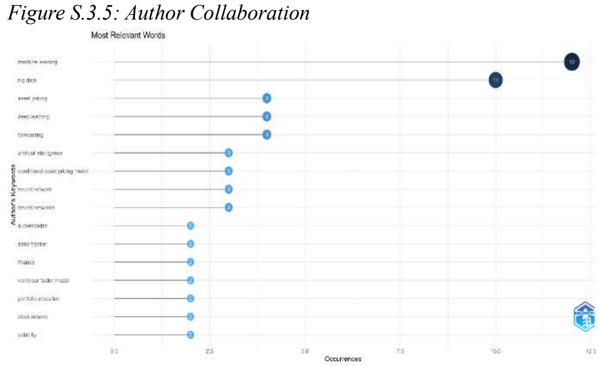


Figure S.3.7: Most Frequent Words

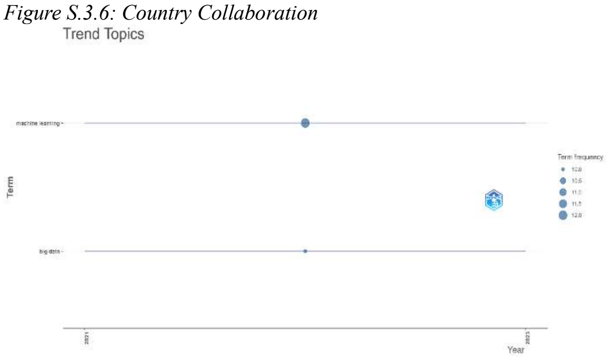


Figure S.3.8: Trend Topics

Note: One author had 2 publications thus, first 10 authors are presented as Most Relevant Authors. Authors' Production over time is a presentation of the most relevant authors and their production over time. 7 countries are captured as corresponding authors countries. The Most Cited countries represent the corresponding author-affiliated countries (Denmark has 0 citations but there is a publication from Switzerland that has 24 publications which is presented in Supplementary Table 5; (Weigand, 2019)). Most Relevant Words and Trend Topics graphs were analysed using Authors' Keywords. Weigand (2019) was not in Web of Science (though the customer service had proved it was in their database); it had to be done manually for other analysis (see Supplementary Table 5) except for the graphs. That is the limitation of the software (VosViewer and Biblioshiny). it does not analyse merged files from different databases.

Table B1: Bibliometric Presentation of Most Productive Countries and Publication Sources

Country	No of Articles	Total Citations	Publication Sources	No of Articles
AI and Asset Pricing				
USA	24	985	Journal of Econometrics	4
China	21	183	Journal of Forecasting	3
Korea	4	1	Review of Financial Studies	3
France	3	12	Economics Letters	2
United Kingdom	3	19	Expert Systems with Applications	2
Australia	2	28	Finance Research Letters	2
Germany	2	10	IEEE Access	2
Italy	2	6	Journal of Computational Science	2
Pakistan	2	10	Journal of Economic Dynamics and Control	2
Spain	2	4	Journal of Intelligent and Fuzzy systems	2
Switzerland	2	25	Journal of Risk and Financial Management	2
Vietnam	2	4	Management Science	2
			Quantitative Finance	2
			Quarterly Journal of Finance	2
Big Data and Asset Pricing				
China	20	181	Journal of Econometrics	13
USA	15	733	Journal of Business and Economic Statistics	4
Spain	4	55	Management Science	3
United Kingdom	3	7	International Journal of Forecasting	2
Brazil	2	27	International Review of Financial Analysis	2
France	2	8	Journal of Financial Econometrics	2
Germany	2	20	Journal of Risk and Financial Management	2
Canada	1	38		
Ireland	1	7		
Switzerland	1	24		
Big Data, AI and Asset pricing				
China	8	19	Journal of Econometrics	2
USA	8	611	Ain Shams Engineering Journal	1
Brazil	2	27	Applied Stochastic Models in Business and Industry	1
Canada	1	38	Connection Science	1
France	1	5	Energy Economics	1
Ireland	1	7	European Financial Management	1
Italy	1	14	European Journal of Finance	1
Switzerland	1	24	Financial Markets and Portfolio Management	1
			International Journal of Financial Studies	1
			International Journal of Forecasting	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 B2: Bibliographic Coupling and Co-citation Analysis of AI and Asset Pricing

Panel A: Bibliographic Coupling

Clusters (*Theme*)	**Author(s) (Year)**
1 (*Factor models and machine learning in asset pricing)*	• Bai and Ng (2019) • Fan et al. (2017) • Giglio et al. (2021) • Giglio et al. (2022) • Gu et al. (2021) • Maehashi and Shintani (2020) • Uddin and Yu (2020)
2 (*AI models and asset pricing theories)*	• Caporale et al. (2016) • Erfanian et al. (2022) • Ho et al. (2011) • Huang and Hueng (2008) • Refenes et al. (1994) • Weller (2018) • Yuan and Lee (2015)
3 (*Deep learning and machine learning in asset pricing*)	• Cao et al. (2005) • Cao et al. (2011) • L. Y. Chen et al. (2024) • Heaton et al. (2017) • Weigand (2019)
4 (*Machine learning and advanced factor models in asset pricing*)	• Jan and Ayub (2019) • Karolyi and Van Nieuwerburgh (2020)
5 (*Asset price dynamics and machine learning*)	• Panchenko et al. (2013) • Warin and Stojkov (2021)

Panel B: Co-Citation Analysis

Clusters (*Theme*)	**Cited Reference**	**Total Citations**
1 (*Factor models, AI and Asset pricing*)	• Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, *52*(1), 57-82.	17
	• Cochrane, J. H. (2011). Presidential address: Discount rates. The Journal of finance, 66(4), 1047-1108.	10
	• Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. Journal of political economy, 81(3), 607-636.	15
	• 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.	17
	• 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.	17
	• 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
	• 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.	14
	• 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
	• Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of financial economics*, *108*(1), 1-28.	10
2 (*Factor analysis and asset pricing theories*)	• 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.	36
	• Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. Journal of financial economics, 116(1), 1-22.	31
	• Lintner, J. (1975). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. In Stochastic optimization models in finance (pp. 131-155). Academic Press.	11
	• 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
	• 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

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 23 papers that share a minimum of 10 shared cited references out of the 81 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 21 cited references meet this threshold out of 2825 cited references.

Table B3: Bibliographic Coupling and Co-citation Analysis of Big Data and Asset Pricing

Bibliographic Coupling

Clusters (*Theme*)	**Author(s) (Year)**
1 (*Machine learning and deep learning in asset pricing and forecasting analysis*)	• L. Y. Chen et al. (2024) • Costa et al. (2021) • Coulombe et al. (2022) • Gu et al. (2021) • Heaton et al. (2017) • Weigand (2019)
2 (*Analysing high-frequency data and large covariance matrices*)	• Dai et al. (2019) • Fan et al. (2016) • Kong and Liu (2018) • Lee and Mykland (2012) • Shephard and Xiu (2017)
3 (*High-frequency data and financial modelling*)	• Ait-Sahalia et al. (2020) • Hollstein et al. (2020) • H. W. Zhang et al. (2023)
4 (*Factor analysis and volatility prediction*)	• He et al. (2022) • Kim and Fan (2019) • Masoliver et al. (2006)

Co-Citation Analysis

Clusters (*Theme*)	**Cited Reference**	**Total Citations**
1 (*Factor-based asset pricing models and theories*)	• 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
	• 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).	13
	• Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, *19*(3), 425-442.	11
2 (*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
	• 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
3 (*High-dimensional data and factor analysis*)	• 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.	14
	• 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 18 papers that share a minimum of 10 cited references out of the 53 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 18 cited references meet this threshold out of 1899 cited references.

Table B4: Bibliographic Coupling and Co-citation Analysis of Big Data, AI and Asset pricing

Bibliographic Coupling		Co-Citation Analysis		
Clusters (*Theme*)	**Author(s) (Year)**	**Clusters (*Theme*)**	**Cited Reference**	**Total Citations**
1 (*Deep learning and machine learning in asset pricing*)	• L. Y. Chen et al. (2024) • Gu et al. (2021) • Heaton et al. (2017) • Karolyi and Van Nieuwerburgh (2020) • Weigand (2019)	1 (*Forecasting analysis and asset pricing theories*)	• Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, *70*(1), 191-221. • 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. • 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). • 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 5 5 5
2 (*Machine learning and forecasting analysis*)	• Costa et al. (2021) • Coulombe et al. (2022) • Kim and Swanson (2018)	2 (*Machine learning and asset pricing*)	• Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273. • Gu, S., Kelly, B., & Xiu, D. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429-450.	11 6
3 (*Advanced models for high-frequency data analysis*)	• Centanni and Minozzo (2006) • Shephard and Xiu (2017)		• 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. • 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.	5 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 9 papers that share a minimum of 10 cited references out of the 24 papers.

b.) The 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 5 citations of a cited reference, only 8 cited references meet this threshold out of 1964 cited references.

Appendix C: List of References of Included Papers

Aït-Sahalia, Y., Kalnina, I., & Xiu, D. C. (2020). High-frequency factor models and regressions. *Journal of Econometrics*, *216*(1), 86-105. https://doi.org/10.1016/j.jeconom.2020.01.007

Akyildirim, E., Nguyen, D. K., Sensoy, A., & Sikic, M. (2023). Forecasting high-frequency excess stock returns via data analytics and machine learning. *European Financial Management*, *29*(1), 22-75. https://doi.org/10.1111/eufm.12345

Alaminos, D., Esteban, I., & Salas, M. B. (2023). Neural networks for estimating Macro Asset Pricing model in football clubs. *Intelligent Systems in Accounting Finance & Management*, *30*(2), 57-75. https://doi.org/10.1002/isaf.1532

Alves, R. P., de Brito, D. S., Medeiros, M. C., & Ribeiro, R. M. (2023). Forecasting Large Realized Covariance Matrices: The Benefits of Factor Models and Shrinkage*. *Journal of Financial Econometrics*, *22*(3), 696-742. https://doi.org/10.1093/jjfinec/nbad013

Andersen, T. G., Riva, R., Thyrsgaard, M., & Todorov, V. (2023). Intraday cross-sectional distributions of systematic risk. *Journal of Econometrics*, *235*(2), 1394-1418. https://doi.org/10.1016/j.jeconom.2022.11.001

Anis, H. T., & Kwon, R. H. (2021). A sparse regression and neural network approach for financial factor modeling. *Applied Soft Computing*, *113*, 15, Article 107983. https://doi.org/10.1016/j.asoc.2021.107983

Atak, A., & Kapetanios, G. (2013). A factor approach to realized volatility forecasting in the presence of finite jumps and cross-sectional correlation in pricing errors. *Economics Letters*, *120*(2), 224-228. https://doi.org/10.1016/j.econlet.2013.03.051

Auh, J. K., & Cho, W. H. (2023). Factor-based portfolio optimization. *Economics Letters*, *228*, 8, Article 111137. https://doi.org/10.1016/j.econlet.2023.111137

Avelar, E. A., & Jordão, R. V. D. (2024). The role of artificial intelligence in the decision-making process: a study on the financial analysis and movement forecasting of the world's largest stock exchanges. *Management Decision*, 24. https://doi.org/10.1108/md-09-2023-1625

Azevedo, V., Kaiser, G. S., & Mueller, S. (2023). Stock market anomalies and machine learning across the globe. *Journal of Asset Management*, *24*(5), 419-441. https://doi.org/10.1057/s41260-023-00318-z

Babikir, A., & Mwambi, H. (2017). Factor Augmented Artificial Neural Network Model. *Neural Processing Letters*, *45*(2), 507-521. https://doi.org/10.1007/s11063-016-9538-6

Bae, J. (2024). Factor-augmented forecasting in big data☆. *International Journal of Forecasting*, *40*(4), 1660-1688. https://doi.org/10.1016/j.ijforecast.2024.02.004

Bagnara, M. (2024). Asset Pricing and Machine Learning: A critical review [Review]. *Journal of Economic Surveys*, *38*(1), 27-56. https://doi.org/10.1111/joes.12532

Bai, J. S., & Ng, S. (2019). Rank regularized estimation of approximate factor models. *Journal of Econometrics*, *212*(1), 78-96. https://doi.org/10.1016/j.jeconom.2019.04.021

Bang, J., Kang, Y., & Ryu, D. (2024). Potential pricing factors in the Korean market. *Finance Research Letters*, *67*, 6, 105946. https://doi.org/10.1016/j.frl.2024.105946

Begušić, S., & Kostanjčar, Z. (2020). Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series. *IEEE Access*, *8*, 164365-164379. https://doi.org/10.1109/access.2020.3021898

Bodilsen, S. T. (2024). Large-Dimensional Portfolio Selection with a High-Frequency-Based Dynamic Factor Model. *Journal of Financial Econometrics*, *23*(2), 27. https://doi.org/10.1093/jjfinec/nbae018

Boubaker, S., Liu, Z. Y., & Mu, Y. H. (2023). Big data analytics and investment. *Technological Forecasting and Social Change*, *194*, 15, 122713. https://doi.org/10.1016/j.techfore.2023.122713

Cakici, N., Fieberg, C., Metko, D., & Zaremba, A. (2023). Machine learning goes global: Cross-sectional return predictability in international stock markets. *Journal of Economic Dynamics & Control*, *155*, 32, 104725. https://doi.org/10.1016/j.jedc.2023.104725

Camacho, M., & Lopez-Buenache, G. (2023). Factor models for large and incomplete data sets with unknown group structure. *International Journal of Forecasting*, *39*(3), 1205-1220. https://doi.org/10.1016/j.ijforecast.2022.05.012

Cao, Q., Leggio, K. B., & Schniederjans, M. J. (2005). A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market. *Computers & Operations Research*, *32*(10), 2499-2512. https://doi.org/10.1016/j.cor.2004.03.015

Cao, Q., Parry, M. E., & Leggio, K. B. (2011). The three-factor model and artificial neural networks: predicting stock price movement in China. *Annals of Operations Research*, *185*(1), 25-44. https://doi.org/10.1007/s10479-009-0618-0

Caporale, G. M., Gil-Alana, L., Plastun, A., & Makarenko, I. (2016). Intraday Anomalies and Market Efficiency: A Trading Robot Analysis. *Computational Economics*, *47*(2), 275-295. https://doi.org/10.1007/s10614-015-9484-9

Casarin, R., Grassi, S., Ravazzolo, F., & van Dijk, H. K. (2023). A flexible predictive density combination for large financial data sets in regular and crisis periods. *Journal of Econometrics*, *237*(2), 12, 105370. https://doi.org/10.1016/j.jeconom.2022.11.004

Centanni, S., & Minozzo, M. (2006). Estimation and filtering by reversible jump MCMC for a doubly stochastic Poisson model for ultra-high-frequency financial data. *Statistical Modelling*, *6*(2), 97-118. https://doi.org/10.1191/1471082X06st112oa

Cerqueti, R., Iovanella, A., Mattera, R., & Storani, S. (2024). Improving the explainability of autoencoder factors for commodities through forecast-based Shapley values. *Scientific Reports*, *14*(1), 13, Article 19622. https://doi.org/10.1038/s41598-024-70342-5

Chen, D. C., Mykland, P. A., & Zhang, L. (2024). Realized regression with asynchronous and noisy high-frequency and high dimensional data☆. *Journal of Econometrics*, *239*(2), 20, 105446. https://doi.org/10.1016/j.jeconom.2023.02.015

Chen, L. Y., Pelger, M., & Zhu, J. (2024). Deep Learning in Asset Pricing. *Management Science*, *70*(2), 38. https://doi.org/10.1287/mnsc.2023.4695

Chen, Y. C., Kuo, S. M., Liu, Y. L., Wu, Z. Q., & Zhang, F. (2022). Improving Returns on Strategy Decisions through Integration of Neural Networks for the Valuation of Asset Pricing: The Case of Taiwanese Stock. *International Journal of Financial Studies*, *10*(4), 15, 99. https://doi.org/10.3390/ijfs10040099

Cheng, M. M., Liao, Y., & Yang, X. Y. (2023). Uniform predictive inference for factor models with instrumental and idiosyncratic betas. *Journal of Econometrics*, *237*(2), 28, 105373. https://doi.org/10.1016/j.jeconom.2022.11.007

Chiu, K. C., & Xu, L. (2004). Arbitrage pricing theory-based Gaussian temporal factor analysis for adaptive portfolio management. *Decision Support Systems*, *37*(4), 485-500. https://doi.org/10.1016/s0167-9236(03)00082-4

Chochola, O., Husková, M., Prásková, Z., & Steinebach, J. G. (2014). Robust monitoring of CAPM portfolio betas II. *Journal of Multivariate Analysis*, *132*, 58-81. https://doi.org/10.1016/j.jmva.2014.07.016

Costa, A. B. R., Ferreira, P. C. G., Gaglianone, W. P., Guillen, O. T. C., Issler, J. V., & Lin, Y. H. (2021). Machine learning and oil price point and density forecasting. *Energy Economics*, *102*, 21, 105494. https://doi.org/10.1016/j.eneco.2021.105494

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920-964. https://doi.org/10.1002/jae.2910

Cueto, J. M., Grané, A., & Cascos, I. (2020). Models for Expected Returns with Statistical Factors. *Journal of Risk and Financial Management*, *13*(12), 17, Article 314. https://doi.org/10.3390/jrfm13120314

Cui, L. Y., Hong, Y. M., Li, Y. X., & Wang, J. H. (2024). A Regularized High-Dimensional Positive Definite Covariance Estimator with High-Frequency Data. *Management Science*, *70*(10), 7242-7264. https://doi.org/10.1287/mnsc.2022.04138

Cuomo, S., Gatta, F., Giampaolo, F., Iorio, C., & Piccialli, F. (2022). An unsupervised learning framework for marketneutral portfolio. *Expert Systems with Applications*, *192*, 13, 116308. https://doi.org/10.1016/j.eswa.2021.116308

Dai, C. X., Lu, K., & Xiu, D. C. (2019). Knowing factors or factor loadings, or neither? Evaluating estimators of large covariance matrices with noisy and asynchronous data. *Journal of Econometrics*, *208*(1), 43-79. https://doi.org/10.1016/j.jeconom.2018.09.005

De Nard, G., Hediger, S., & Leippold, M. (2022). Subsampled factor models for asset pricing: The rise of Vasa. *Journal of Forecasting*, *41*(6), 1217-1247. https://doi.org/10.1002/for.2859

Diallo, B., Bagudu, A., & Zhang, Q. (2023). Fama-French three versus five, which model is better? A machine learning approach. *Journal of Forecasting*, *42*(6), 1461-1475. https://doi.org/10.1002/for.2970

Dixon, M., & Polson, N. (2020). Short Communication: Deep Fundamental Factor Models. *Siam Journal on Financial Mathematics*, *11*(3), SC26-SC37. https://doi.org/10.1137/20m1330518

Du, H. C. (2023). Research on Amazon's stock price forecasting based on arbitrage pricing model based on big data. *Ain Shams Engineering Journal*, *14*(1), 6, 101793. https://doi.org/10.1016/j.asej.2022.101793

Duppati, G., Sonono, E., Musvoto, W., Scrimgeour, F., & Tiwari, A. K. (2023). Is Liquidity Risk Priced in Artificial Intelligence ETFs? *Australasian Accounting Business and Finance Journal*, *17*(3), 62-80. https://doi.org/10.14453/aabfj.v17i3.05

Erfanian, S., Zhou, Y. W., Razzaq, A., Abbas, A., Safeer, A. A., & Li, T. (2022). Predicting Bitcoin (BTC) Price in the Context of Economic Theories: A Machine Learning Approach. *Entropy*, *24*(10), 29, 1487. https://doi.org/10.3390/e24101487

Escribano, A., Peña, D., & Ruiz, E. (2021). 30 years of cointegration and dynamic factor models forecasting and its future with big data: Editorial [Editorial Material]. *International Journal of Forecasting*, *37*(4), 1333-1337. https://doi.org/10.1016/j.ijforecast.2021.06.004

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. https://doi.org/10.1080/07350015.2015.1052458

Fan, J. Q., Xue, L. Z., & Yao, J. W. (2017). Sufficient forecasting using factor models. *Journal of Econometrics*, *201*(2), 292-306. https://doi.org/10.1016/j.jeconom.2017.08.009

Fang, M., & Taylor, S. (2021). A machine learning based asset pricing factor model comparison on anomaly portfolios. *Economics Letters*, *204*, 7, 109919. https://doi.org/10.1016/j.econlet.2021.109919

Fang, Y., Chen, Y. Z., & Ren, H. (2023). A factor pricing model based on machine learning algorithm. *International Review of Economics & Finance*, *88*, 280-297. https://doi.org/10.1016/j.iref.2023.06.012

Feng, G. H., He, J. Y., Polson, N. G., & Xu, J. N. (2024). Deep Learning in Characteristics-Sorted Factor Models. *Journal of Financial and Quantitative Analysis*, *59*(7), 3001-3036. https://doi.org/10.1017/s0022109023000893

French, J. (2017). The time traveller's CAPM. *Investment Analysts Journal*, *46*(2), 81-96. https://doi.org/10.1080/10293523.2016.1255469

Gao, Y. Y., He, D., Mu, Y., & Zhao, H. M. (2023). Realised volatility prediction of high-frequency data with jumps based on machine learning. *Connection Science*, *35*(1), 16, Article 2210265. https://doi.org/10.1080/09540091.2023.2210265

Ge, S. Y., Li, S. R., & Linton, O. (2023). News-implied linkages and local dependency in the equity market. *Journal of Econometrics*, *235*(2), 779-815. https://doi.org/10.1016/j.jeconom.2022.07.004

Giglio, S., Kelly, B., & Xiu, D. C. (2022). Factor Models, Machine Learning, and Asset Pricing. *Annual Review of Financial Economics*, *14*, 337-368. https://doi.org/10.1146/annurev-financial-101521-104735

Giglio, S., Liao, Y., & Xiu, D. C. (2021). Thousands of Alpha Tests. *Review of Financial Studies*, *34*(7), 3456-3496. https://doi.org/10.1093/rfs/hhaa111

Gu, S. H., Kelly, B., & Xiu, D. C. (2021). Autoencoder asset pricing models. *Journal of Econometrics*, *222*(1), 429-450. https://doi.org/10.1016/j.jeconom.2020.07.009

Han, Y. (2021). Research on influencing factors of stock returns based on multiple regression and artificial intelligence model. *Journal of Intelligent & Fuzzy Systems*, *40*(4), 6457-6467. https://doi.org/10.3233/jifs-189485

He, A., & Zhou, G. F. (2023). Diagnostics for asset pricing models. *Financial Management*, *52*(4), 617-642. https://doi.org/10.1111/fima.12436

He, Y., Kong, X. B., Yu, L., & Zhang, X. S. (2022). Large-Dimensional Factor Analysis Without Moment Constraints. *Journal of Business & Economic Statistics*, *40*(1), 302-312. https://doi.org/10.1080/07350015.2020.1811101

Healy, J. V., Gregoriou, A., & Hudson, R. (2024). Automated Machine Learning and Asset Pricing. *Risks*, *12*(9), 12, Article 148. https://doi.org/10.3390/risks12090148

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. https://doi.org/10.1002/asmb.2209

Ho, W. R. J., Tsai, C. L., Tzeng, G. H., & Fang, S. K. (2011). Combined DEMATEL technique with a novel MCDM model for exploring portfolio selection based on CAPM. *Expert Systems with Applications*, *38*(1), 16-25. https://doi.org/10.1016/j.eswa.2010.05.058

Hollstein, F., Prokopczuk, M., & Simen, C. W. (2020). The Conditional Capital Asset Pricing Model Revisited: Evidence from High-Frequency Betas. *Management Science*, *66*(6), 2474-2494. https://doi.org/10.1287/mnsc.2019.3317

Huang, J. W., Xing, R., & Li, Q. (2022). Asset pricing via deep graph learning to incorporate heterogeneous predictors. *International Journal of Intelligent Systems*, *37*(11), 8462-8489. https://doi.org/10.1002/int.22950

Huang, P., & Hueng, C. J. (2008). Conditional risk-return relationship in a time-varying beta model. *Quantitative Finance*, *8*(4), 381-390. https://doi.org/10.1080/14697680701191361

Huang, W. G. (2022). Digesting Three-factor Model. *Singapore Economic Review*, 30. https://doi.org/10.1142/s0217590822480022

Jan, M. N., & Ayub, U. (2019). Do the Fama and French five-factor model forecast well using ANN? *Journal of Business Economics and Management*, *20*(1), 168-191. https://doi.org/10.3846/jbem.2019.8250

Karolyi, G. A., & Van Nieuwerburgh, S. (2020). New Methods for the Cross-Section of Returns. *Review of Financial Studies*, *33*(5), 1879-1890. https://doi.org/10.1093/rfs/hhaa019

Khan, U. E., & Iqbal, J. (2021). The Relationship between Default Risk and Asset Pricing: Empirical Evidence from Pakistan. *Journal of Asian Finance Economics and Business*, *8*(3), 717-729. https://doi.org/10.13106/jafeb.2021.vol8.no3.0717

Khoa, B. T., & Huynh, T. T. (2022). Long Short-Term Memory Recurrent Neural Network for Predicting the Return of Rate Underframe the Fama-French 5 Factor. *Discrete Dynamics in Nature and Society*, *2022*, 8, 3936122. https://doi.org/10.1155/2022/3936122

Khoa, B. T., & Huynh, T. T. (2023). A comparison of CAPM and Fama French three-factor model under machine learning approaching. *Journal of Eastern European and Central Asian Research*, *10*(7), 1100-1111. https://doi.org/10.15549/jeecar.v10i7.1402

Kiliç, A., Güloglu, B., Yalçin, A., & Üstündag, A. (2023). Big data-enabled sign prediction for Borsa Istanbul intraday equity prices. *Borsa Istanbul Review*, *23*, S38-S52. https://doi.org/10.1016/j.bir.2023.08.005

Kim, D., & Fan, J. Q. (2019). Factor GARCH-Ito models for high-frequency data with application to large volatility matrix prediction. *Journal of Econometrics*, *208*(2), 395-417. https://doi.org/10.1016/j.jeconom.2018.10.003

Kim, E., Cho, T., Koo, B., & Kang, H. G. (2023). Conditional autoencoder asset pricing models for the Korean stock market. *Plos One*, *18*(7), 30, e0281783. https://doi.org/10.1371/journal.pone.0281783

Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339-354. https://doi.org/10.1016/j.ijforecast.2016.02.012

Kong, X. B., & Liu, C. (2018). Testing against constant factor loading matrix with large panel high-frequency data. *Journal of Econometrics*, *204*(2), 301-319. https://doi.org/10.1016/j.jeconom.2018.03.001

Kong, X. B., Liu, Z., & Zhou, W. (2019). A rank test for the number of factors with high-frequency data. *Journal of Econometrics*, *211*(2), 439-460. https://doi.org/10.1016/j.jeconom.2019.03.004

Lalwani, V., & Meshram, V. V. (2022). The cross-section of Indian stock returns: evidence using machine learning. *Applied Economics*, *54*(16), 1814-1828. https://doi.org/10.1080/00036846.2021.1982132

Lee, S. S., & Mykland, P. A. (2012). Jumps in equilibrium prices and market microstructure noise. *Journal of Econometrics*, *168*(2), 396-406. https://doi.org/10.1016/j.jeconom.2012.03.001

Levin, P. (2023). Instantaneous and Averaged Volatility in Two-Side Filtration Model of Financial Asset Pricing. *Journal of Derivatives*, *31*(1), 49-63. https://doi.org/10.3905/jod.2023.1.182

Li, J., Todorov, V., & Tauchen, G. (2019). Jump factor models in large cross-sections. *Quantitative Economics*, *10*(2), 419-456. https://doi.org/10.3982/qe1060

Li, W. P., & Mei, F. (2020). Asset returns in deep learning methods: An empirical analysis on SSE 50 and CSI 300. *Research in International Business and Finance*, *54*, 21, Article 101291. https://doi.org/10.1016/j.ribaf.2020.101291

Liu, Z. Y., Teka, H., & You, R. Y. (2023). Conditional autoencoder pricing model for energy commodities. *Resources Policy*, *86*, 10, 104060. https://doi.org/10.1016/j.resourpol.2023.104060

Lo, A. W., & Singh, M. (2023). Deep-learning models for forecasting financial risk premia and their interpretations. *Quantitative Finance*, *23*(6), 917-929. https://doi.org/10.1080/14697688.2023.2203844

Loo, W. K. (2020). Predictability of HK-REITs returns using artificial neural network. *Journal of Property Investment & Finance*, *38*(4), 291-307. https://doi.org/10.1108/jpif-07-2019-0090

Luo, A., Zhong, L., Wang, J. L., Wang, Y., Li, S. J., & Tai, W. P. (2024). Short-Term Stock Correlation Forecasting Based on CNN-BiLSTM Enhanced by Attention Mechanism. *IEEE Access*, *12*, 29617-29632. https://doi.org/10.1109/access.2024.3369419

Ma, T., Leong, W. J., & Jiang, F. W. (2023). A latent factor model for the Chinese stock market. *International Review of Financial Analysis*, *87*, 21, 102555. https://doi.org/10.1016/j.irfa.2023.102555

Ma, T., Liao, C. F., & Jiang, F. W. (2023). Timing the factor zoo via deep learning: Evidence from China. *Accounting and Finance*, *63*(1), 485-505. https://doi.org/10.1111/acfi.13033

Maasoumi, E., Wang, J. Q., Wang, Z., & Wu, K. (2024). Identifying factors via automatic debiased machine learning. *Journal of Applied Econometrics*, *39*(3), 438-461. https://doi.org/10.1002/jae.3031

Maehashi, K., & Shintani, M. (2020). Macroeconomic forecasting using factor models and machine learning: an application to Japan. *Journal of the Japanese and International Economies*, *58*, 17, 101104. https://doi.org/10.1016/j.jjie.2020.101104

Martínez, R. G., Casado, P. P., & Román, M. P. (2020). Market efficiency analysis using Al models based on Investors' Mood. *Revista Perspectiva Empresarial*, *7*(2), 10-23. https://doi.org/10.16967/23898186.649

Masoliver, J., Montero, M., Perelló, J., & Weiss, G. H. (2006). The continuous time random walk formalism in financial markets. *Journal of Economic Behavior & Organization*, *61*(4), 577-598. https://doi.org/10.1016/j.jebo.2004.07.015

Massa Roldán, R., Reyna Miranda, M., & Gómez Salcido, V. (2022). Neuro-wavelet Model for price prediction in high-frequency data in the Mexican Stock market [Modelo de neuro-onda para predicción de precios en datos de alta frecuencia en el Mercado Bursátil Mexicano]. *Revista mexicana de economía y finanzas*, *17*(1), e570. https://doi.org/10.21919/remef.v17i1.570

Meng, C., Chen, C., Xu, H., & Li, T. (2024). Asset Pricing and Portfolio Investment Management Using Machine Learning: Research Trend Analysis Using Scientometrics. *Economics-the Open Access Open-Assessment E-Journal*, *18*(1), 20, Article 20220108. https://doi.org/10.1515/econ-2022-0108

O'neill, W. D. (1995). Rate-distortion theory and the volatility of asset prices. *Information Sciences-Applications*, *3*(1), 55-74. https://doi.org/10.1016/1069-0115(95)82004-3

Oomen, R. C. A. (2010). High-dimensional covariance forecasting for short intra-day horizons. *Quantitative Finance*, *10*(10), 1173-1185, 920959412. https://doi.org/10.1080/14697680903220349

Pan, S. Y., Long, S. W., Wang, Y. M., & Xie, Y. (2023). Nonlinear asset pricing in Chinese stock market: A deep learning approach. *International Review of Financial Analysis*, *87*, 9, 102627. https://doi.org/10.1016/j.irfa.2023.102627

Panchenko, V., Gerasymchuk, S., & Pavlov, O. V. (2013). Asset price dynamics with heterogeneous beliefs and local network interactions. *Journal of Economic Dynamics & Control*, *37*(12), 2623-2642. https://doi.org/10.1016/j.jedc.2013.06.015

Park, W., Song, Y. U., Chun, D., & Kim, J. (2024). The five-factor model analysed by machine learning classification techniques. *Applied Economics Letters*, 7. https://doi.org/10.1080/13504851.2024.2308576

Peterburgsky, S. (2021). Is aggregate volatility a priced risk factor? *International Review of Finance*, *21*(3), 843-864. https://doi.org/10.1111/irfi.12299

Refenes, A. N., Zapranis, A., & Francis, G. (1994). Stock performance modeling using neural networks - A comparative-study with regression-models. *Neural Networks*, *7*(2), 375-388. https://doi.org/10.1016/0893-6080(94)90030-2

Rudkin, W. (2020). Mispricing, returns and the quest for parsimony. *Finance Research Letters*, *37*, 5, Article 101368. https://doi.org/10.1016/j.frl.2019.101368

Rui, X. Y. (2023). Influencing factors of stock returns based on Fama-French model and intelligent algorithm. *Soft Computing*, *27*(14), 10359-10368. https://doi.org/10.1007/s00500-023-08305-7

Sanhaji, B., & Chevallier, J. (2023). Tracking 'Pure' Systematic Risk with Realized Betas for Bitcoin and Ethereum. *Econometrics*, *11*(3), 36, Article 19. https://doi.org/10.3390/econometrics11030019

Shang, M. S., Luo, X., Liu, Z. G., Chen, J., Yuan, Y., & Zhou, M. C. (2019). Randomized Latent Factor Model for High-dimensional and Sparse Matrices from Industrial Applications. *Ieee-Caa Journal of Automatica Sinica*, *6*(1), 131-141. https://doi.org/10.1109/jas.2018.7511189

Shephard, N., & Xiu, D. C. (2017). Econometric analysis of multivariate realised QML: Estimation of the covariation of equity prices under asynchronous trading. *Journal of Econometrics*, *201*(1), 19-42. https://doi.org/10.1016/j.jeconom.2017.04.003

Sun, Y. C., & Xu, W. (2022). A Factor-Based Estimation of Integrated Covariance Matrix with Noisy High-Frequency Data. *Journal of Business & Economic Statistics*, *40*(2), 770-784. https://doi.org/10.1080/07350015.2020.1868301

Sun, Y. C., Xu, W., & Zhang, C. H. (2023). Identifying latent factors based on high-frequency data. *Journal of Econometrics*, *233*(1), 251-270. https://doi.org/10.1016/j.jeconom.2022.04.006

Tang, C., Yu, Q. C., Li, X. N., Lu, Z. K., & Yang, Y. F. (2023). Application of DeepForest-CQP multi-factor model in quantitative stock selection. *Journal of Intelligent & Fuzzy Systems*, *44*(3), 5425-5436. https://doi.org/10.3233/jifs-222328

Uddin, A., Tao, X. Y., & Yu, D. T. (2023). Attention based dynamic graph neural network for asset pricing. *Global Finance Journal*, *58*, 25, 100900. https://doi.org/10.1016/j.gfj.2023.100900

Uddin, A., & Yu, D. T. (2020). Latent factor model for asset pricing. *Journal of Behavioral and Experimental Finance*, *27*, 11, 100353. https://doi.org/10.1016/j.jbef.2020.100353

Wang, C. (2024). Stock return prediction with multiple measures using neural network models. *Financial Innovation*, *10*(1), 34, 72. https://doi.org/10.1186/s40854-023-00608-w

Wang, Q. Y. (2021). Cryptocurrencies asset pricing via machine learning. *International Journal of Data Science and Analytics*, *12*(2), 175-183. https://doi.org/10.1007/s41060-021-00252-6

Warin, T., & Stojkov, A. (2021). Machine Learning in Finance: A Metadata-Based Systematic Review of the Literature [Review]. *Journal of Risk and Financial Management*, *14*(7), 31, 302. https://doi.org/10.3390/jrfm14070302

Weigand, A. (2019). Machine learning in empirical asset pricing [Article]. *Financial Markets and Portfolio Management*, *33*(1), 93-104. https://doi.org/10.1007/s11408-019-00326-3

Weller, B. M. (2018). Does Algorithmic Trading Reduce Information Acquisition? *Review of Financial Studies*, *31*(6), 2184-2226. https://doi.org/10.1093/rfs/hhx137

Yang, J. H., & Ma, M. J. (2022). The numerical solution of one-dimensional discrete asset pricing model based on the improved trigonometric extreme learning machine. *Journal of Computational Science*, *63*, 8, 101809. https://doi.org/10.1016/j.jocs.2022.101809

Yang, J. H., Zhang, M., Feng, S., Zhang, X. J., & Bai, X. (2024). A hierarchical deep model integrating economic facts for stock movement prediction. *Engineering Applications of Artificial Intelligence*, *133*, 14, 108320. https://doi.org/10.1016/j.engappai.2024.108320

Yang, X. L., Zhu, Z. F., Li, D., & Zhu, K. (2024). Asset Pricing via the Conditional Quantile Variational Autoencoder. *Journal of Business & Economic Statistics*, *42*(2), 681-694. https://doi.org/10.1080/07350015.2023.2223683

Yao, H. X., Xia, S. H., & Liu, H. (2022). Six-factor asset pricing and portfolio investment via deep learning: Evidence from Chinese stock market. *Pacific-Basin Finance Journal*, *76*, 28, Article 101886. https://doi.org/10.1016/j.pacfin.2022.101886

Yao, H. X., Xia, S. H., & Liu, H. (2024). Return predictability via an long short-term memory-based cross-section factor model: Evidence from Chinese stock market. *Journal of Forecasting*, *43*(6), 1770-1794. https://doi.org/10.1002/for.3096

Yuan, F. C., & Lee, C. H. (2015). Using least square support vector regression with genetic algorithm to forecast beta systematic risk. *Journal of Computational Science*, *11*, 26-33. https://doi.org/10.1016/j.jocs.2015.08.004

Zapata, H. O., & Mukhopadhyay, S. (2022). A Bibliometric Analysis of Machine Learning Econometrics in Asset Pricing. *Journal of Risk and Financial Management*, *15*(11), 17, 535. https://doi.org/10.3390/jrfm15110535

Zeng, Q. D., Bing, T., Li, L., & Xu, Y. (2024). Data Factor and Financial Market Equilibrium. *Emerging Markets Finance and Trade*, *60*(4), 663-677. https://doi.org/10.1080/1540496x.2023.2251652

Zhan, B. Q., Zhang, S., Du, H. S., & Yang, X. G. (2022). Exploring Statistical Arbitrage Opportunities Using Machine Learning Strategy. *Computational Economics*, *60*(3), 861-882. https://doi.org/10.1007/s10614-021-10169-8

Zhang, H. W., Zhao, X. Y., Gao, W., & Niu, Z. B. (2023). The role of higher moments in predicting China's oil futures volatility: Evidence from machine learning models. *Journal of Commodity Markets*, *32*, 27, 100352. https://doi.org/10.1016/j.jcomm.2023.100352

Zhang, R. X., Zhao, C. Y., & Lin, G. L. (2023). Interpretable image-based deep learning for price trend prediction in ETF markets. *European Journal of Finance*, 29. https://doi.org/10.1080/1351847x.2023.2275567

Zhang, X., Yang, X., Li, J. P., & Hao, J. (2023). Contemporaneous and noncontemporaneous idiosyncratic risk spillovers in commodity futures markets: A novel network topology approach. *Journal of Futures Markets*, *43*(6), 705-733. https://doi.org/10.1002/fut.22407

Zhu, L., Basu, S., Jarrow, R. A., & Wells, M. T. (2020). High-Dimensional Estimation, Basis Assets, and the Adaptive Multi-Factor Model. *Quarterly Journal of Finance*, *10*(4), 52, 2050017. https://doi.org/10.1142/s2010139220500172

Zhu, L., Jarrow, R. A., & Wells, M. T. (2021). Time-Invariance Coefficients Tests with the Adaptive Multi-Factor Model. *Quarterly Journal of Finance*, *11*(04), 30, 2150019. https://doi.org/10.1142/s2010139221500191