Highlights

Financial Time Series Uncertainty: A Review of Probabilistic AI Applications

Sivert Eggen, Tord Johan Espe, Kristoffer Grude, Morten Risstad, Rickard Sandberg

- This paper surveys the rapidly growing, but still immature, literature on probabilistic AI in financial time-series uncertainty forecasting.
- We find that the potential for financial decision-making provided by probabilistic AI, originating from their inherent capabilities to distinguish epistemic from aleatoric uncertainty, remains largely underutilized. This is despite growing regulatory concerns and the well-documented challenges of stability, interpretability, trustworthiness, accountability, and risk management in many traditional machine learning applications.
- The current body of literature is characterized by a lack of standardized benchmarks and evaluation metrics, limited replicability, and insufficient financial interpretation of results, which hinders clear conclusions regarding the applicability of probabilistic AI models in real-world applications.
- We argue that interdisciplinary research is needed to advance the field.

Financial Time Series Uncertainty: A Review of Probabilistic AI Applications

Sivert Eggen[a], Tord Johan Espe[a], Kristoffer Grude[a], Morten Risstad[a,*,1] and Rickard Sandberg[b]

[a]*Norwegian University of Science and Technology, Institute of Industrial Economics and Technology Management, Alfred Getz v. 4, 7004, Trondheim, Norway*

[b]*Stockholm School of Economics, Department of Entrepreneurship, Innovation and Technology, Stockholm, Sweden*

ARTICLE INFO

Keywords:
Machine learning
Probabilistic AI
Time series predictions
Epistemic uncertainty
Aleatoric uncertainty
Tail risk

ABSTRACT

Probabilistic machine learning models offer a distinct advantage over traditional deterministic approaches by quantifying both epistemic uncertainty (stemming from limited data or model knowledge) and aleatoric uncertainty (due to inherent randomness in the data), along with full distributional forecasts. These capacities are particularly appealing in light of growing regulatory concerns and the well-documented challenges of stability, interpretability, trustworthiness, accountability and risk management in many machine learning applications.

This review of probabilistic artificial intelligence in financial time-series uncertainty forecasting highlights several critical gaps in the field. These include a lack of standardized benchmarks and evaluation metrics, limited interdisciplinary collaboration, and insufficient financial interpretation of results. Collectively, these shortcomings hinder the ability to draw definitive conclusions about the performance of probabilistic models. The field remains nascent and fragmented, with most research published only recently and few studies building upon prior work — likely due in part to the infrequent disclosure of code.

We conclude that the potential for financial decision-making provided by probabilistic AI remains largely underutilized.

1. Introduction

The use of artificial intelligence (AI)[1] and machine learning (ML) models in finance has seen a significant recent surge, arguably motivated by their inherent capabilities of handling a high-dimensional target and feature space with potentially non-linear dynamics and correlated predictors. Despite this increasing interest in AI-driven approaches, traditional econometric approaches, such as ARIMA (AutoRegressive Integrated Moving Average; Box and Jenkins (1970)) and GARCH (Generalized AutoRegressive Conditional Heteroskedasticity; Bollerslev (1986)), still prevail in the domain of financial time series modeling (López de Prado, 2019). These models have demonstrated potential in capturing the complexities of financial markets, but have primarily been used to provide point forecasts rather than full conditional probability distributions (Tang, Song, Zhu, Yuan, Hou, Ji, Tang and Li, 2022). Beyond financial forecasting, probabilistic AI models offer significant implications for business decision-making. The ability to quantify and distinguish between epistemic and aleatoric uncertainty enables financial managers to refine capital allocation strategies, optimize risk-adjusted returns, and enhance strategic planning under uncertainty. For instance, corporations can leverage uncertainty-aware AI models in dynamic pricing, supply chain financing, and investment portfolio adjustments. Furthermore, financial institutions integrating uncertainty estimation into credit risk assessment can improve lending decisions by dynamically adjusting risk premiums based on market conditions. Moreover, uncertainty-aware AI models can mitigate algorithmic bias in credit scoring and automated trading by identifying high-risk predictions before execution. By bridging AI-driven forecasting with corporate financial strategies, probabilistic AI provides an actionable framework for data-driven business decisions. The aim of this survey is to review the scientific literature on applications of probabilistic AI models to enhance uncertainty estimates in financial time series.

*Corresponding author

morten.risstad@ntnu.no (M. Risstad)

ORCID(s): 0000-0003-2562-8892 (M. Risstad); 0000-0003-0589-4034 (R. Sandberg)

[1]See Appendix A for a list of all abbreviations used in this paper.

Furthermore, we propose directions for future research to facilitate probabilistic AI models as catalysts for improved financial decision-making.[2]

When quantifying prediction uncertainty, the distinction between epistemic (model-driven) and aleatoric (underlying) uncertainty is important. The former arises from the model's limitations in capturing data patterns, caused by not having enough data, not using the relevant features, or model misspecification. This type of uncertainty is reducible through improving models or data. Conversely, the latter refers to inherent data randomness, equivalent to latent volatility in finance. This is influenced by unpredictable market behavior, economic events, and investor sentiment, beyond the reach of any model. The total uncertainty is thus the sum of the epistemic variance and the aleatoric variance. A sophisticated probabilistic AI model is able to quantify both types of uncertainty. This has profound implications for the potential implementation of ML models in the finance industry, from the perspectives of both practitioners and regulators. A major drawback of AI models is that they are to a large extent "black boxes". This complicates interpretations of model outputs and the underlying determinants. Additionally, ML models are sensitive to training procedures, parameter tuning and data quality. Hence, output from deterministic ML models are liable to be highly sensitive to small perturbations in input data and parameters. This inherent model risk limits practical applications. Probabilistic models, on the other hand, are capable of outputting prediction intervals, accompanied by parameter uncertainty estimates. This is extremely useful in practical applications, for different reasons. First and foremost, it is well suited to address concerns put forward by regulators with respect to interpretability and trustworthiness of AI models.[3] Investors in efficient markets can benefit from improved understanding of risk dynamics. For instance, an investor might construct a portfolio to maximize returns relative to aleatoric uncertainty, while simultaneously imposing a threshold to exclude predictions where epistemic uncertainty is unacceptably high. In sum, probabilistic AI models provide a versatile and flexible framework for analyzing financial time series, ultimately facilitating more informed, adaptive, and robust investment decisions.

Several recent reviews on applications of AI and machine learning for financial time series prediction are available. Gandhmal and Kumar (2019), Li and Bastos (2020) and Kumbure, Lohrmann, Luukka and Porras (2022) review machine learning techniques applied to stock market trend or point predictions, up to 2018, 2019 and 2019 respectively. Gandhmal and Kumar (2019) conclude that Artificial Neural Networks (ANNs) and fuzzy-based techniques are the most promising among the reviewed machine learning approaches for accurate stock market predictions, supported by Shi and Zhuang (2019) review of soft computing approaches, finding ANN architectures to consistently outperform other machine learning models in point prediction accuracy. Li and Bastos (2020) and Sezer, Gudelek and Ozbayoglu (2020) show that RNN based models like Long-Short Term Memory (LSTM) implementations are the most popular in deep learning. Kumbure et al. (2022) conclude that the most frequently utilized models are ANNs and Support Vector Machines (SVM), but that deep learning models like LSTM have growing interest due to reports of robust and improved predictions. Khattak, Shafi, Khan, Flores, Lara, Samad and Ashraf (2023) provides an in-depth review of machine learning methods applied to forecast various financial assets between 2018 and 2023 and find new hybrid integrations of LSTM and SVM architectures to be more effective than traditional models for point predictions. Gunnarsson, Isern, Kaloudis, Risstad, Vigdel and Westgaard (2024) survey the relevance of AI models for volatility predictions, and report promising results across asset classes.

Common to the aforementioned studies is their focus on deterministic models and point predictions. Even though both epistemic and aleatoric uncertainty are essential components of financial decision-making, few reviews have been conducted on the topic of uncertainty quantification in a financial context using machine learning methods or probabilistic models. A notable exception is Abdar et al. (2021), who conduct a review on uncertainty quantification using deep learning techniques. They discuss advantages and disadvantages of several models but do not focus on financial time series predictions. The authors conclude that Deep Ensembles and Bayesian Neural Networks show promising capabilities for uncertainty quantification distinguishing between aleatoric and epistemic uncertainty, applicable to financial forecasting, but find that lack of standardized benchmarks make it difficult to compare frameworks. The study closest to ours is Blasco, Sánchez and García (2024), who conduct a survey on uncertainty quantification using deep learning techniques in financial time series. Blasco et al. (2024) show that most papers

[2]Probabilistic AI is a rapidly evolving field, and includes Bayesian neural networks, Variational Autoencoders, Gaussian processes, and Probabilistic Neural Networks among others. Broadly speaking, we include any machine learning models that can quantify the predictive uncertainty or produce full distributional forecasts.

[3]For instance, the Basel Committee calls for tail risk models that allow for stress testing based on well-defined risk factors (Basel Committee on Banking Supervision, 2019). Probabilistic AI models are capable of both identifying relevant risk factors while also assessing their relative importance, accounting for both epistemic and aleatoric uncertainty, and hence improve robustness, interpretability and trustworthiness.

do not distinguish between aleatoric and epistemic uncertainty, and few authors perform analysis on the financial implications of predictive uncertainty. The review is limited to deep learning models using a Bayesian approach and investigates methods for approximation of posterior distributions in these models. Their focus on how to use and interpret uncertainty estimates in a financial context is limited, and they do not assess how researchers evaluate uncertainty estimates or discuss the appropriate way of doing so. Specifically, they do not assess whether probabilistic AI models truly are an improvement over econometric models and traditional AI models when it comes to their practical utility in financial applications and decision-making. We distinguish ourselves from Blasco et al. (2024) in at least three important aspects. First, we use a broader definition of probabilistic AI and hence capture a broader span of relevant models. Second, we expand the scope and analyze a higher number of relevant dimensions; including model output, asset classes, and type of uncertainty. Third, while their survey includes literature on deep learning up to 2022, we use data up until 2024. This is important, in light of the rapid growth in the body of scientific literature in this field over the last two years.

This review addresses a significant gap in the literature by systematically analyzing the application of probabilistic AI models in finance, with a focus on their potential to improve uncertainty estimation in financial forecasting. While existing literature reviews largely focus on AI from a machine learning perspective, our work connects AI-driven probabilistic modeling with real-world financial challenges, including market volatility, systemic risk, portfolio optimization, and financial regulation. Detailed analysis is contained in the remainder of the paper, still some overall conclusions emerge. While probabilistic AI is a rapidly evolving field, its adoption in financial practice remains limited due to concerns over interpretability, computational complexity, and integration into existing decision-making frameworks. The literature employing probabilistic AI still predominantly focuses on point predictions and largely does not exploit the potential for quantification of aleatoric nor epistemic uncertainty. We find that empirical investigations of predictive performance generally lack appropriate benchmarking and rigorous evaluation of statistical and economic model performance. Furthermore, we note that the majority of active researchers have a computer science background, as opposed to finance or economics. This leaves a potential for inter-disciplinary research collaboration to advance the field, through employing state-of-the-art probabilistic AI models within a rigorous scientific framework and robust empirical applications.

Beyond summarizing existing models, our review provides a roadmap for future research by highlighting key methodological gaps, regulatory considerations, and the need for interdisciplinary collaboration between AI researchers, economists, and financial practitioners. The remaining sections of this paper is organized as follows: Section 2 covers Data and Methodology, detailing the review and analysis process. Section 3 presents the Results along what we consider to be the most important dimensions of the data. Section 4 discusses important aspects of these results and summarizes potential avenues for further research, while Section 5 concludes. Appendix A lists abbreviations, Appendix B categorizes journals, Appendix C offers a detailed description of the models discussed, and Appendix D presents key attributes of the sampled papers.

2. Data and Methodology

To ensure reproducible and unbiased results we adopt a structured systematic literature review (SLR) approach adhering to Snyder (2019) and Marzi, Balzano, Caputo and Pellegrini (2024).[4]

To comprehensively capture relevant literature, we utilize multiple well-established databases. SCOPUS and Web of Science were chosen due to their extensive coverage and comprehensive indexing of academic literature within a wide range of fields, as well as being the most extensively used databases for reviews (Marzi et al., 2024). IEEE Xplore was included being a leading database for review papers in the fields of computer science and engineering (Suhaimi and Abas, 2020; Carvalho, Soares, Vita, Francisco, Basto and Alcalá, 2019; Cavacini, 2015), while ProQuest is another large academic database utilized by Gunnarsson et al. (2024) and others.

We design search criteria to ensure that as many papers relevant to the research questions as possible are included in the initial search, as a narrow search query in this phase can lead to involuntary exclusion of relevant documents (Marzi et al., 2024; Kuhrmann, Méndez Fernández and Daneva, 2017; Williams Jr, Clark, Clark and Raffo, 2021). By requiring papers to match at least one term in four different clauses, we ensure that every paper was (1) within the field of AI, (2) about probabilistic modeling, (3) about forecasting, and (4) within finance. The keywords used are inspired

[4]More specifically, we (1) define research questions and boundaries, (2) define search queries, (3) select databases, (4) screen and cross-check data and (5) clean and export data. All relevant code, including detailed database search queries, are available from the corresponding author upon request.

Table 1
Keywords used in database search queries across four key areas.

Category	Keywords
(1) Artificial Intelligence (AI)	`AI, ML, Artificial intelligence, Machine learning, Deep learning, Reinforcement learning, Supervised learning`
(2) Probabilistic Modeling	`Probabilistic, Uncertainty quantification, Prediction interval*, Confidence interval*, Distributional forecast, Bayesian, Gaussian process, Undirected graphical model*, Markov Network*, Markov random field*, Probabilistic Graphical Model*, Variational inference, Monte Carlo dropout, Hidden Markov model*, Gaussian mixture model*, Variational Autoencoder*, Dirichlet Process`
(3) Forecasting	`Forecast*, Predict*, Estimat*`
(4) Finance	`Foreign exchange, Forex, Equity market*, Stock price*, Stock market*, Stock return*, Stock trend*, Stock index, Stock indices, Commodities, Value-at-risk, Value at risk, CVaR, Expected shortfall, Financial time series, Implied volatility, Realized volatility, Cryptocurrency, Bitcoin, (Volatility AND Financ*)`

* The asterisk (*) is a wildcard character.
Papers must match at least one term in each of the categories to be included in the initial sample.

by Blasco et al. (2024), but we restructured the query to ensure that all articles meet each of the four criteria defining our scope, which is slightly broader. Table 1 shows all keywords included within each clause. Only peer-reviewed scientific papers in the English language are included.

Following the initial search, we employ a set of inclusion criteria in a two-stage process. These criteria ensure that the paper discusses a model that predicts the price of some financial instrument, the model must be an AI or machine learning model, and the model must be able to provide predictive output beyond point predictions. Figure 1 illustrates the cleaning and screening process. The final sample consists of 62 articles published between 2004 and 2024.

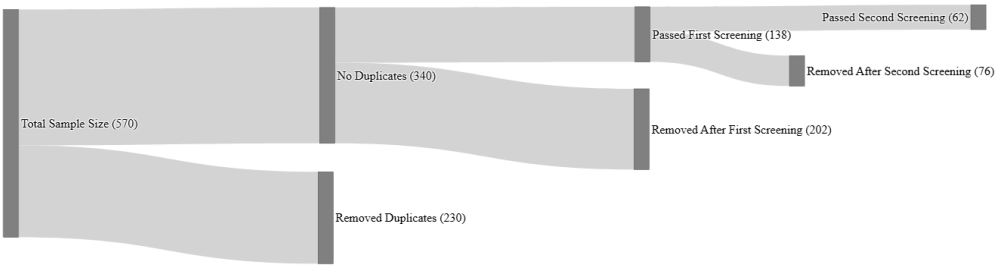

Figure 1: Flow chart illustrating the reduction in sample size throughout cleaning and screening phases.

Scientific output in the field has accelerated in recent years, with the majority of papers published after 2020, peaking in 2024 with 13 publications, as illustrated in Figure 2.

Notably, significant contributions to the field are made by authors[5] from computer science and technical faculties, while only 16% originate from researchers within finance, business or economics, as illustrated in Figure 3. In terms of journal origins, the majority of papers are published in engineering, technical, computer science, and artificial intelligence journals or others. Finance and economics journals have limited representation, with 15 out of 62 papers. Figure 4 shows the distribution of publications across journal categories. Only 11 of 62 papers disclose the code for the proposed models. This lack of disclosure constitutes a significant obstacle for reproducibility, as the models are often too complex to enable reproduction based written exposition only. Furthermore, this hinders scientific advances in general since it becomes difficult for researchers to build on existing research.

Figure 5 illustrates the specific financial assets and markets that are forecasted in the studies. The majority of papers focus on equities, while a notable number of papers also address currencies and derivatives.

[5]To capture the full scope of expertise, all authors are counted individually.

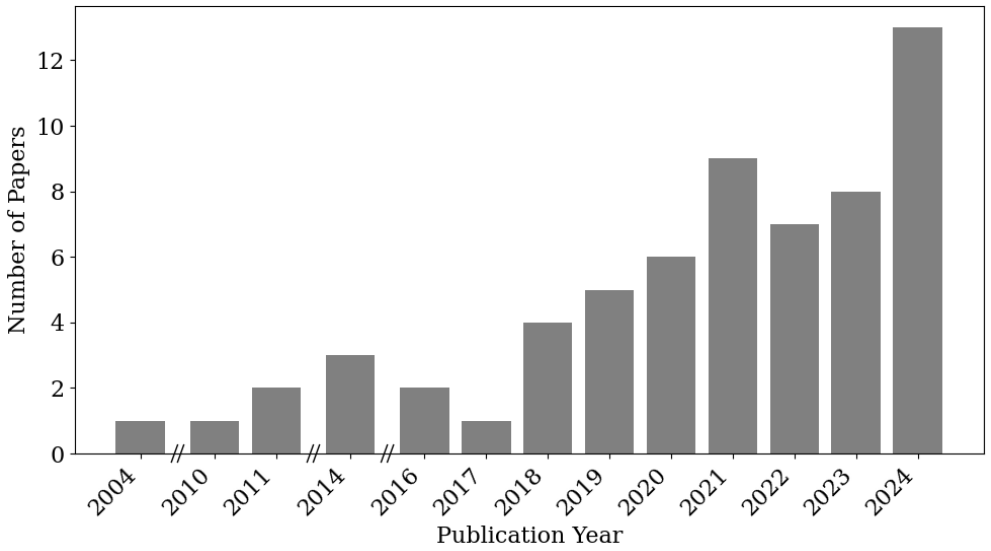

Figure 2: Annual distribution of papers in the field published included in the sample, illustrating the recent increase in publications.

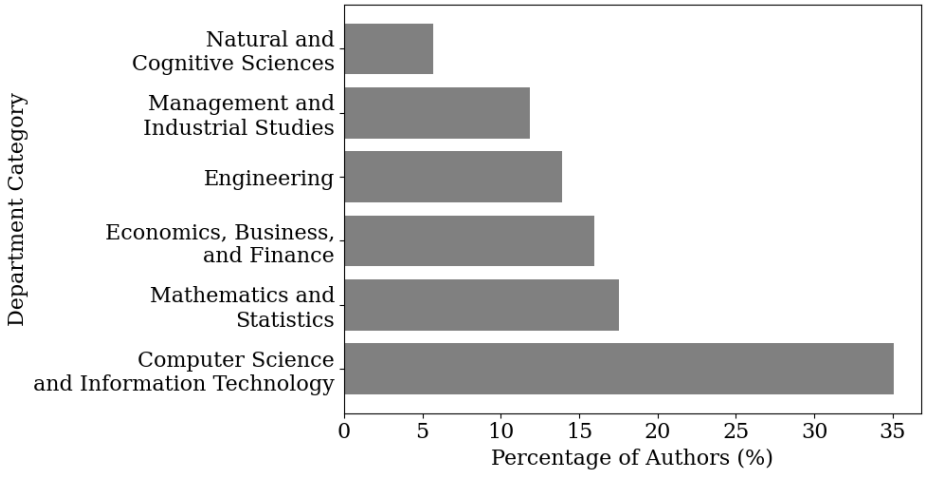

Figure 3: Percentage distribution of authors by academic faculty category.

Notes: The categorization of journals is detailed in Appendix B. To capture the full scope of expertise, all authors are counted individually.

3. Results

This section presents the results along with what we consider to be the most important dimensions of the dataset. Note that a descriptive table summarizing key attributes for each of the included papers is available in Appendix D. In Section 3.1 we describe applications of the most predominantly used probabilistic models and provide an overview of how they are used to create uncertainty estimates for financial time series. We avoid technical detailed expositions, and refer to Appendix C for a brief overview of the models. Section 3.2 categorizes the sample based on model outputs. Motivation for making predictions and quantifying uncertainty, as well as the feasibility of doing so, vary across asset classes. In Section 3.3 we provide an overview of how probabilistic AI has been applied to different asset classes, including challenges encountered in each specific case. All models in the sample are capable of providing predictions and related uncertainty quantifications. We outline the extent to which this is utilized in Section 3.4.

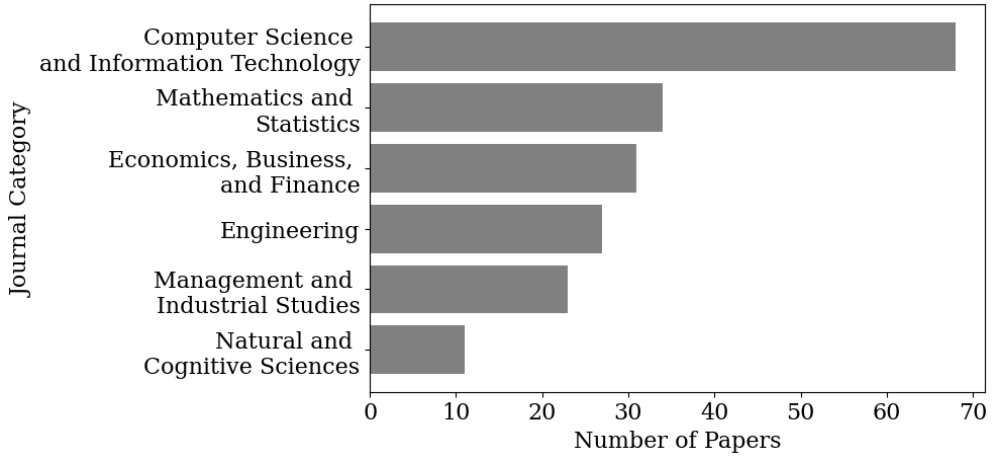

Figure 4: Distribution of sample papers by journal category.

Notes: The categorization of journals is detailed in Appendix B.

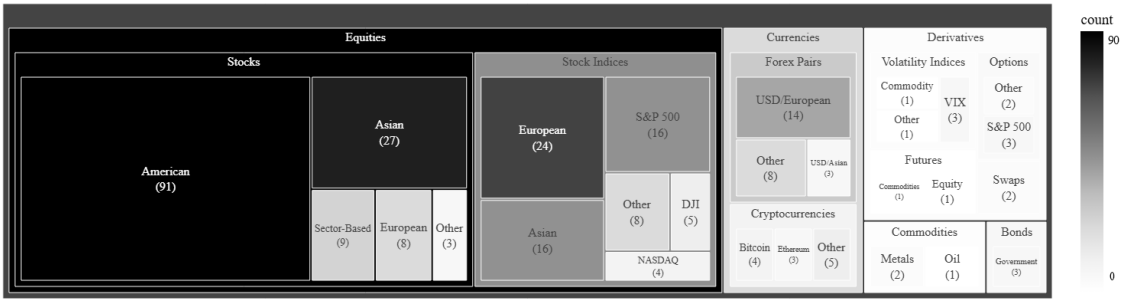

Figure 5: Distribution of financial markets and assets targeted by the predictive models in the sample papers.

Notes: Each individual asset predicted in every paper is counted once.

3.1. Models

Table 2 contains an overview of probabilistic model categories from grouping the most commonly used models. Figure 6 illustrates the occurrence of each probabilistic model category, and whether the model is used independently or in combination with other machine learning or econometric models.

3.1.1. Bayesian Neural Networks (BNNs).

Cocco, Tonelli and Marchesi (2021) and Jang and Lee (2018a) employ BNNs for cryptocurrency price predictions, primarily focusing on point prediction accuracy. Cocco et al. (2021) apply a BNN with Monte Carlo approximation to predict daily Bitcoin and Ethereum prices, benchmarking against LSTM and Feed Forward Neural Networks (FFNNs). The BNN underperforms on Bitcoin in terms of MAPE but yield better results for Ethereum, while deployed two-stage models with a Support Vector Regression (SVR) and LSTM or FFNN outperform in all cases. Although the authors use the BNN's outputted quantiles as prediction confidence, the authors are focused on point predictions and do not assess uncertainty. Similarly, Jang and Lee (2018a) employ a BNN to make point predictions for Bitcoin price and volatility, using blockchain-specific data, outperforming linear regression and SVR on MAPE and RSME. The authors present confidence intervals for price and volatility, which could be used to assess total uncertainty. However, the probabilistic output is not leveraged to integrate these measures, nor is uncertainty explicitly evaluated as focus lie on point prediction accuracy. Notably, predictions frequently exceed the stated upper and lower bounds.

Chandra and He (2021) apply a BNN with Markov Chain Monte Carlo (MCMC) for multi-step stock price forecasting, benchmarking against FFNNs trained with ADAM and SGD. The BNN provide superior point estimates in terms of RSME for all stocks. The authors use the probabilistic output of the BNN to create prediction intervals as

Table 2
Probabilistic Model Categorization.

Model Category	Models
Bayesian Neural Networks (BNN)	BNN, Gen-BNN, B-TABL
Gaussian Processes Regression (GPR)	GP, GPR, G4P, GPMCH
Variational Autoencoders (VAE)	VAE
Hidden Markov Models (HMM)	HMM, CHMM, MCHMM
Probabilistic Recurrent Neural Network (P-RNN) Extensions	DeepAR, DeepARA, P-GRU, QRBiLSTM, ESVM, Bayesian LSTM, Bayes ES-RNN, Clockwork RNN
Probabilistic Generative Adversial Networks (P-GAN)	cGAN, PredACGAN
Probabilistic Neural Networks (PNN)	PNN
Other Bayesian Methods	B-SVR, BGLM, Bayesian Network, MCMC
Other Probabilistic AI Methods	RSMAN, Recurrent Dictionary Learning (RDL), TV-Entropy, Probabilistic Fuzzy Logic (PFL), Leave-One-Out Cross-Conformal Predictive System (LOO-CCPS), P-SVM, B-HANN, Probabilistic Graphical Model (PGM), Augmented DCNN

a measure of uncertainty. However, the quality or robustness of the estimate is not assessed, and evidently the actual stock price frequently fall outside the bounds for some stocks, indicating an unreliable uncertainty estimate. The authors compare uncertainty levels during and after Covid, showing higher predicted uncertainty during the pandemic.

Soleymani and Paquet (2022) propose a hybrid model, QuantumPath, combining a BNN with a temporal GAN to predict long-term prices for several S&P 500 stocks. The BNN predicts the drift and volatility parameters for a Feynman-Dirac integral, which simulate stock trajectories by Monte Carlo, while the temporal GAN generates trajectories by condsidering the most probable paths. The probabilistic BNN output is used to estimate the underlying probability distribution of the stock trajectories, and is therefore used implicitly as a volatility estimate. The models weighted expected values for 30-day predictions outperform ARIMA and Ornstein-Uhlbeck. Even though the trajectories represent a distribution of prices, the uncertainty is not assessed.

Hortúa and Mora-Valencia (2024) employ a Bayesian Neural Network (BNN) to forecast the VIX using a hybrid architecture that integrates WaveNet, a Temporal Convolutional Network (TCN), with Bayesian inference techniques such as the Reparametrization Trick (RT), Flipout, and Multiplicative Normalizing Flows (MNF). The authors apply quantile recalibration to correct the miscalibration tendency in neural networks, addressing potentially unreliable uncertainty estimates due to error overestimation, by aligning observed and expected data proportions within prediction intervals, assessed using Root Mean Squared Calibration Error (RMSCE). The models using MNF demonstrate the most calibrated predictions, and generally superior short-term point predictions compared to ARIMA.

Magris, Shabani and Iosifidis (2023) introduce a Bayesian Temporal Augmented Bilinear Neural Network (B-TABL) for forecasting and classifying mid-price changes in Limit Order Books (LOB). Employing a Variational Online Gauss Newton (VOGN) method for Bayesian inference, the model yields better calibrated class probabilities than approaches like Monte Carlo Dropout. Expected Calibration Error (ECE) and Expected Calibration Distance (ECD) are used to evaluate how well predicted probabilities align with actual observed frequencies, assessing model reliability in uncertainty estimation. The BNN framework provides predictive distributions for class probabilities, offering an uncertainty measure the authors interpret as confidence. While VOGN optimizer for B-TABL does not clearly outperform ADAM on standard classification metrics, the authors argue that the model deliver more meaningful classifications due to superior calibration scores.

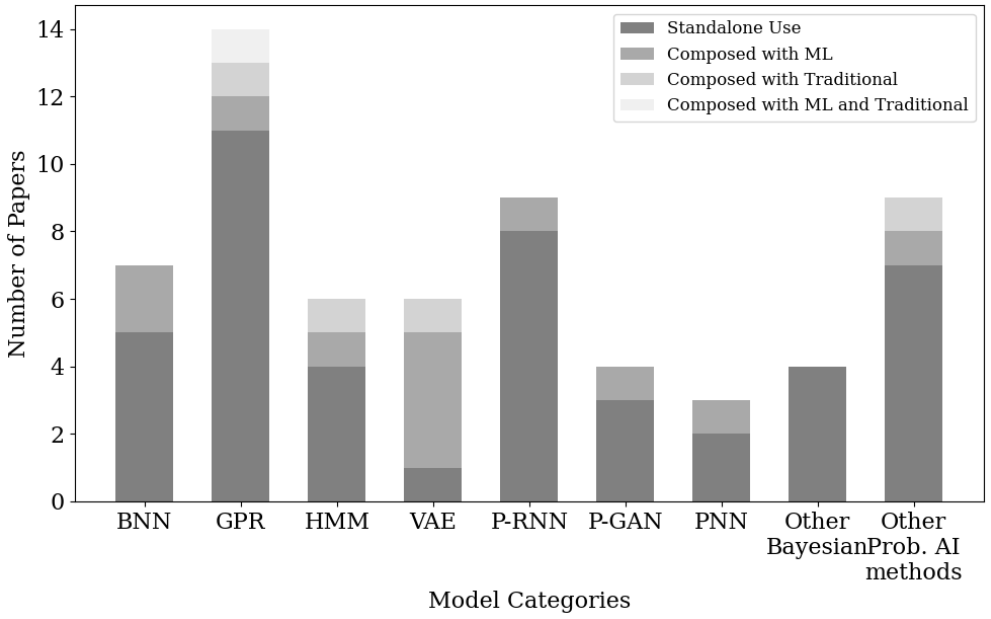

Figure 6: Breakdown of probabilistic model usage in research papers across standalone and hybrid approaches.

Lastly, Jang and Lee (2018b) use a Generative BNN to predict American put option prices of differing moneyness on the S&P 500 by incorporating prior data from financial models. By extending a standard BNN with domain-specific priors and self-evolving capabilities, the model accounts for data-scarce regions like deep in-the-money or out-of-the-money options. The authors solely focus on point predictions of the price and do not quantify uncertainty in predictions. Compared to standard BNN implementations, a GPR model, and a GARCH implementation, their model is superior in terms of MAPE and RMSE.

3.1.2. Gaussian Process Regression (GPR).

Suphawan, Kardkasem and Chaisee (2022) employ GPR to forecast the Stock Exchange of Thailand (SET). The model is compared to ANN and RNN and demonstrate superior prediction accuracy on traditional error metrics (RMSE, MAE, MAPE, NSE). The authors note that the distributional output makes the GPR model advantageous compared to ANN and RNN models because it allows for prediction results with quantification of uncertainty, but the quality of this uncertainty is not assessed.

Wang, Feng, Li, He and Feng (2021b) incorporate GPR in a multi-scale nonlinear ensemble model to predict price with uncertainty for the S&P 500, Dow Jones and NASDAQ. The ensemble model use Variational Mode Decomposition (VDM) and an Autoencoder (AE) for feature extraction, and a two-step deep learning setup with RNN and LSTM. The GPR is used in final stage to create interval predictions and uncertainty estimates. The model is benchmarked against a regular GPR, ANN, RNN and LSTM implementations, displaying superiority in MAPE, MSE, RMSE, MAE and SSE for point predictions. Additionally, the interval predictions are assessed on coverage probability metrics like Mean Width Percentage (MWP), Mean Coverage (MC) and Prediction Interval Coverage Probability (PICP), outperforming a standalone GPR. (Wang, Feng, Li, He and Feng, 2021a) present an alternative ensemble (SSA-EWSVM-RNN-GPR) for stock index forecasting. Singular Spectrum Analysis (SSA) is used to decompose the original stock index signal to several components for preprocessing and feature extraction, and Enhanced Weighted Support Vector Machine (EWSVM) forecast the decomposed components, which are then combined with a RNN to make point forecasts. Finally, point predictions are fed into a GPR to generate interval forecasts. MWP and CP are used to validate the interval forecast, scoring better than eight GPR benchmark models. Improved point prediction accuracy is also reported.

Li, Xia, Li, Kuruoğlu, Jiang and Xia (2024b) integrate GPR with a graph-aware portfolio selection model with Generalized Gaussian Distribution (GGD) likelihood. The model captures both mean return and variance, used for portfolio selection, but do not account for heteroskedacity in the modeling. The GPR-model's selected portfolios return

is compared to traditional methods like UCRP, PAMR and OLMAR, yielding better performance in terms of annualized return, Sharpe ratio, annualized volatility and max drawdown on stock data from NYSE, S&P and TSE. Platanios and Chatzis (2014) propose a Gaussian Process Mixture Conditional Heteroscedasticity (GPMCH) model to forecast price and volatility for currency exchange rates and global large-cap equity indices. The model captures volatility clustering and handle the non-linear dependency, presenting a viable alterative to traditional GARCH models. The authors applies a Pitman-Yor process to better capture skewed and tail heavy data distributions, showing that their their GPMCH outperformed GARCH in volatility predictions. Tegnér and Roberts (2021) use a GPR to transform market option price data into a smooth implied volatility surface, capturing implied volatility across options of various strike prices and maturities. Then values of this surface is forecasted using GPR, enabling predictions of implied volatility and, consequently, future option prices. The results indicate a promising ability to forecast the VIX one week ahead, outperforming a naive forecasting approach, but no other benchmark models are considered.

Risk and Ludkovski (2018) use a GPR to construct portfolio tail risk measures in VaR and TVaR. Monte Carlo simulation is used to generate training data for the GPR model and to evaluate portfolio losses under diverse economic scenarios. The GPR model employs non-parametric spatial modeling, meaning it does not assume a fixed functional form for the relationship between economic scenarios and portfolio losses. Instead, it adapts dynamically based on the data, leveraging the principle that similar economic conditions tend to produce similar portfolio outcomes. This approach reduces bias and variance in risk estimates compared to nested simulations. In one of the advanced models presented, they use heteroskedastic GP (hetGP), which handles scenarios of varying levels of noise more effectively, further enhancing the uncertainty estimates. Similarly, Li, Li, Min and Lin (2023) use a GPR to predict return distributions of several American technology stocks, used to derive investor confidence levels based on prediction standard deviations, integrated in a Black-Litterman framework for portfolio construction. The method is benchmarked against an equal weighting and a Global Minimum Variance (GMV) method, and is not able to clearly outperform the GMV in Sharpe ratio or maximum drawdown.

The remaining studies in the sample utilizing GPR—Papaioannou, Talmon, Kevrekidis and Siettos (2022); Žmuk and Jošić (2020); Park, Kim and Lee (2014); Hendawy, McMillan, Sakr and Shahwan (2023) and Spiegeleer, Madan, Reyners and Schoutens (2018)—apply GPR for price or return prediction in financial time series. Höcht, Schoutens and Verschueren (2024) predict forward-looking implied volatility (IVOL), generally demonstrating competitive performance against benchmarked models. However, these works focus primarily on point predictions without assessing the probabilistic outputs of the GPR model or quantifying prediction uncertainty.

3.1.3. Variational Autoencoders (VAEs).

One single paper applies a VAE independently. Arian, Moghimi, Tabatabaei and Zamani (2022) propose Encoded VaR, directly applying VAE to estimate VaR by generating synthetic market scenarios from historical cross-sectional stock returns of the S&P 500, LSE and FSE. The VAE learns the latent structure of the financial return distributions without relying on parametric assumptions or predefined joint distributions, and in turn generate samples of synthetic returns to be interpreted as potential future outcomes. The VAE architecture allows generation of arbitrarily many samples, enabling reconstruction of a theoretical underlying distribution for VaR calculation. The authors claim to enhance the signal-to-noise ratio present in financial data, and benchmark the VaR estimate against GARCH models. While the model shows competitive results for specific loss functions like Lopez' method (Lopez, 1998), it does not pass all adequacy tests and is not valid, in contrast to GARCH extensions CaViaR-GARCH and EVT-GARCH.

Of the papers utilizing VAEs in combination with other models, several use it as a probabilistic input to another model and do not directly infer uncertainty estimates from the probabilistic output of the VAE. Caprioli, Cagliero and Crupi (2023) apply a VAE for risk management to assess credit portfolio sensitivity to asset correlations. The VAE is used to generate synthetic correlation matrices, simulating various market conditions, used as input in a multi-factor Vasciek model with Monte Carlo simulation to examine how shifts in correlations affect VaR. Choudhury, Abrishami, Turek and Kumar (2020) use VAEs as a pre-processing tool to denoise NASDAQ stock financial time series before using a stacked LSTM autoencoder to make point predictions. The authors report superior results in point predictions compared to other machine learning models like, but do not assess uncertainty in their forecasts. Tang, Huang and Rinprasertmeechai (2024) also deploy VAEs for denoising financial time series data by extracting latent representations. The model is combined with a transformer (LPAST), and is used for long-term multi-step point predictions of different financial times series. The proposed method outperforms benchmarked machine learning models in point predictions, but the probabilistic outputs of the VAE are not directly utilized for uncertainty quantification. Li, Cui, Wang, Liu, Qin and Yang (2020) combine a multimodal VAE with a LSTM architecture to

predict agriculture commodity futures. The VAE is used to extract high-level features and reduce noise for input data in the LSTM, and probabilistic output is not used specifically in predictions. Their proposed model outperforms ARIMA, and machine learning benchmarks like CNNs in point prediction accuracy.

Xing, Cambria and Zhang (2019) propose an innovative approach by combining VAEs with a RNN to forecast stock volatility. Their model, Sentiment-Aware Volatility Forecasting (SAVING), integrates social media sentiment data to jointly model stock price movements and the sentiment that influences them. This interaction is captured through the VAE's latent variables, from which marginal joint probabilities are inferred. Benchmarked against econometric models GARCH, EGARCH and TARCH, the SAVING model outperforms in terms of negative log-likelihood (NLL).

3.1.4. Hidden Markov Models (HMMs).

Two articles use HMMs to address multi-asset dependencies in financial time series forecasting. Li and Cheng (2010) propose a stochastic HMM for forecasting fuzzy time series data, modeling the Taiwan Weighted Stock Index as the hidden states and the New Taiwan dollar against the U.S. dollar as the observable state. While the model performs better compared to a standard HMM implementation in forecasting accuracy, it is not evaluated against any other models. Additionally, focus lie solely on point predictions, without assessing the probabilistic output of the model. Cao, Zhu and Demazeau (2019) extend the multi-factor dependency approach by developing a Multi-Layer Coupled HMM (MCHMM). Unlike Li and Cheng (2010) who address dependencies within a single market, Cao et al. (2019) capture interactions both within and between markets, specifically between stock and currency markets across different countries. The model is reportedly more accurate than ARIMA and logistic regression in trend prediction of German and Dutch stock markets. However, similar to Li and Cheng (2010), uncertainty in predictions derivable from the probabilistic model output is not assessed.

In Park, Lee and Lee (2011), historical segments of financial time series are labeled as either up-trending or down-trending using the Perceptually Important Points (PIP) algorithm. Continuous Hidden Markov Models (CHMMs) are subsequently trained to classify out-of-sample data. The results demonstrate that the HMM model significantly outperform Support Vector Machines (SVMs) across most tested assets, including currencies, stock indices, and individual stocks.

Sher, Rehman, Kim and Ihsan (2023) also forecast categorical return trends, applying HMM alongside several other models to forecast movements in individual technology stocks. Although the details around their specific model implementation are limited, the authors report superior performance from the HMM compared to ARIMA, LSTM and several booster models. The probabilistic output of HMM is leveraged to assess the likelihood of future stock price movements, but like the previous studies, uncertainty assessment is not addressed.

Zhang, Jiang, Fang, Zeng and Xu (2019) extend the HMM to a second-order model, capturing both short-term and long-term dependencies for predicting next-day categorical trends in stock indices. In this higher order approach, the observation depends not only on current state, but also on previous hidden states. While the authors do not directly use the probabilistic distribution output to assess uncertainty, they suggest that the higher-order HMM has lower risk then the first-order model, supported by improved Sharpe ratios and reduced maximum drawdown in their trading strategy experiment. Additionally, the second-order HMM deliver better predictive performance compared to the first-order HMM. Similarly, Su and Yi (2022) apply the second-order HMM model to predict prices and directions of the Hang Seng Index (HSI). The authors focus exclusively on point predictions and do not assess uncertainty of any kind. Compared to NA-GARCH, CNN-BiLSTM-AM and AHMMAS, the second-order HMM showcase superior performance in RMSE and MAE.

3.1.5. Probabilistic RNN Extensions (P-RNN).

Several articles apply Bayesian methods within RNN implementations, imposing prior distributions over the network weights to estimate the posterior distribution in Eq.(5). Hassan (2024) utilize a Bayesian LSTM with MC dropout at inference, optimized with ADAM, to generate a distributional forecast of Bitcoin prices. The model outperforms non-Bayesian LSTMs in RMSE, R^2 and MAPE for point predictions, and the Bayesian approach facilitates epistemic uncertainty quantification. The author argues that the model uncertainty is accurately estimated, as it increases with prediction distance from actual data, but no other assessment measure of the quality of the uncertainty estimate is utilized. Similarly, Dixon (2022) incorporate exponential smoothing within a Bayesian RNN, smoothing hidden states to capture long-term dependencies in IBM stock price predictions. The model provides more accurate forecasts compared to a standard LSTM and GRU implementation, with better coverage of confidence intervals across various predictive horizons. This improvement is presented as evidence of superior uncertainty estimates.

Parker, Holan and Wills (2021) present a Bayesian General Bayesian Heteroskedacity Model (GBHM) within a RNN framework to predict Dow Jones index volatility. Compared to a GARCH implementation, the model achieves superior log predictive scores. Additionally, the authors report more accurate uncertainty measured by coverage, with GARCH yielding an inflated 100% coverage for the 50% prediction interval, while the model attains nearly optimal 50%. Previous research has shown that GARCH generally produces reliable coverage probabilities when modeling stock indices (Rippel and Jánský, 2011), raising questions about whether the GARCH model benchmarked against is misspecified. Tian, Niu and Wei (2023) forecast volatility indices using a Clockwork RNN optimized with a Multi-Objective Grey Wolf optimizer, employing empirical mode decomposition to capture both linear and non-linear trends. The model produces deterministic and probabilistic forecasts, with uncertainty quantified and assessed using PICP, PINAW, and Winkler score. It demonstrates superior accuracy in point predictions and stability across case studies compared to ARIMA and LSTM implementations, including the VIX, crude oil ETF volatility index (COEVI), and the 10-year U.S. treasury note volatility index (TYVIX). Uncertainty estimates are not benchmarked against other models. Golnari, Komeili and Azizi (2024) introduce a probabilistic GRU model incorporating Bayesian inference to treat network weights as probabilistic, enabling distributional forecasts for cryptocurrency price predictions. The model is superior to LSTM and GRU implementations in MAPE and R^2 in point predictions. The authors use the standard deviation of the forecasted distributions as a measure of prediction uncertainty, but do not further assess it's reliability or distinguish uncertainty types. Wang and Lin (2024) employ Quantile Regression (QR) within a Bi-Directional LSTM model to produce probabilistic range predictions for gold prices, incorporating several macroeconomic factors. The QR-BiLSTM predicts multiple quantiles of the future price distribution, capturing price fluctuations and is used as a measure of uncertainty. The authors assess the total uncertainty of the predicted distributions, without separating model and underlying uncertainty, using the Average Internal Score (AIS), which balances interval width and accuracy. The model outperform other LSTM and GRU benchmarks on this metric.

Three articles in the sample employ the DeepAR model. (Salinas, Flunkert and Gasthaus, 2019) use an autoregressive RNN-based model that generates parameters of a predefined probability distribution at each time step. Fatouros, Makridis, Kotios, Soldatos, Filippakis and Kyriazis (2023) apply DeepAR to forecast VaR for a forex portfolio, comparing it to GARCH and other models using Christoffersen's and Dynamic Quantile tests for adequacy. Promising results are reported as the adequacy tests are passed, and superior accuracy in most loss functions is achieved. Almeida, Müller and Perlin (2024) use DeepAR to forecast VaR and ES for crypto liquidity pool portfolios, reporting superior accuracy in ES prediction over GARCH. However, without any adequacy tests, interpretation of results is unvalid. Li, Chen, Zhou, Yang and Zeng (2024a) extend DeepAR with an attention mechanism (DeepARA) for stock price forecasting in the Chinese market, achieving superior MAPE in point predictions compared to other neural networks. The authors assess uncertainty by analyzing the entropy of the predicted price distributions, concluding that the model provides good estimates, but lack comparative uncertainty evaluation, as no alternative models considered provided comparable distributions.

3.1.6. Probabilistic Generative Adversarial Networks

Four articles use Probabilistic GANs for financial time series forecasting. Lee and Seok (2021) propose a modified conditional GAN (cGAN-UC) to forecast the price of NASDAQ-100 Future Index. The generator is used to produce outputs based on multiple different sampled noise vectors combined with input features, generating a range of outputs for the same input, forming a distribution of predictions. The model outperform deterministic model implementations (ANNs and Random forests) in point prediction accuracy, and the quality of uncertainty estimates, assessed by how well the estimated uncertainty correlates with actual prediction errors, is superior compared to a standard BNN implementation.

Vuletić, Prenzel and Cucuringu (2024) introduce a specialized GAN model (Fin-GAN) for one-step-ahead probabilistic forecasting of stock and ETF return distributions. The model employs a custom loss function for the GAN's generator, emphasizing the directional accuracy of forecasts while integrating uncertainty. Using a basic long-short trading strategy based on the signs of forecasts and incorporating uncertainty-weighted trade sizes for the Fin-GAN model only, the model outperforms ARIMA and LSTM benchmarks following the same strategy measured by Sharpe ratio and variance in PnL. The quality of the produced uncertainty estimates is not assessed outside the trading context. Similarly, Kim and Lee (2023) employ a predictive auxiliary classifier GAN (PredACGAN) for portfolio optimization, incorporating prediction uncertainty in S&P 500 and NASDAQ 100 stocks. The generator forecasts future return distributions, classifying stocks as long, short or hold, while a portfolio is constructed and rebalanced monthly by combining expected return with the entropy of distribution as a risk measure to maximize risk-adjusted

returns. Compared with a uniform portfolio and portfolios based on the predictions of other models (MLP and gradient boosting) without applying risk measures, PredACGAN incorporating uncertainty demonstrates superior performance in terms of Sharpe ratio and max drawdown, indicating the uncertainty estimates were meaningful.

Salama (2024) apply a cGAN model integrated with a spotted hyena optimization algorithm for hyperparameter tuning to forecast stock prices. Compared to GAN-based implementations with alternative tuning strategies, the model achieves superior accuracy in terms of MAE and MSE for predicted price. Although the model generates a full probabilistic distribution, the author neither assesses nor utilize it for uncertainty estimation.

3.1.7. Probabilistic Neural Networks (PNNs).

Thawornwong and Enke (2004) utilize a PNN to predict the directions of future excess stock return for a portfolio of stocks listed on the S&P 500. Adaptive selection of economic variables for prediction using recent relevant variables is performed. The model outperform models such as linear regression, random walk and neural networks with constant variables. The authors focus on directional accuracy and risk-adjusted profits to provide a trading strategy with reduced risk and increased profitability, rather than uncertainty quantification assessment. Chandrasekara, Tilakaratne and Mammadov (2019) enhance the standard PNN by introducing a multivariate scaled t-distribution as the joint distribution of input variables to capture the heavy-tailed and correlated nature of financial data, addressing the limitations of the Gaussian assumption commonly used in PNNs. To address the multi-class imbalance problem, a multi-class undersampling based bagging (MCUB) technique is proposed, balancing class distributions and improving classification accuracy. Tested on three stock indices (AORD, GSPC, and ASPI), superior directional accuracy over standard PNN models is demonstrated, but the paper does not emphasize uncertainty quantification in these classifications. Lahmiri (2011) compare a PNN with a Back-Propagation Neural Network that is optimized using Genetic algorithms (GA-BPNN), for predicting daily trends in tech stocks and the NYSE index. GA-BPNN outperforms the traditional PNN in accuracy, and the authors do not assess the calibration of the probabilities generated by either model.

3.1.8. Other Bayesian Methods.

Other Bayesian Methods refer to models applying Bayesian techniques and are able to quantify uncertainty, without using Neural Network architectures. Rather than delving into the technical implementations of each model, we will focus on summarizing the key results they achieve.

Malagrino, Roman and Monteiro (2018) utilize a Bayesian network to predict the directional movements of the iBOVESPA index by incorporating dependencies among multiple global stock indices, achieving competitive accuracy with comparable literature. The authors classify binary without explicitly quantifying uncertainty in classification outcomes. Similarly, Raúl, Plaza Casado and Prado Román (2021) apply a Bayesian network to forecast IBEX index trends, incorporating investor sentiment to enhance model performance. The authors interpret the classification probabilities as trust levels that indicate degrees of uncertainty, and develop a trading system shown to systematically outperform the market. Grudniewicz and Slepaczuk (2023) evaluate a Bayesian Generalized Linear Model (BGLM) alongside traditional and machine learning models to classify stock movements to generate trading signals across various indices for algorithmic trading. Their findings indicate that algorithmic trading outperform passive strategies, with BGLM being among the most accurate models. However, the probabilistic output of the BGLM is not use to assess uncertainty, nor integrated into the trading strategies. A distinct application is proposed by Law and Shawe-Taylor (2017), using Bayesian Support Vector Regression (B-SVR) for price prediction and prediction uncertainty estimates for various financial time series, including equity indices, commodity futures and bond yields. The Bayesian framework optimizes model parameters, and the model produces interval predictions. The authors evaluate the uncertainty estimate quality by examining the correlation between prediction uncertainty and actual errors using the Coefficient of Variation (CoV), classifying predictions as reliable or unreliable based on a continuously calibrated threshold value, excluding unreliable predictions. The model is not benchmarked against traditional or ML models.

3.1.9. Other Probabilistic AI Methods.

Other Probabilistic AI models refer to methods that are capable of producing probabilistic forecasts, but do not fit easily in any of the aforementioned categories.

Daniali, Barykin, Kapustina, Khortabi, Sergeev, Kalinina, Mikhaylov, Veynberg, Zasova and Senjyu (2021) employ a Deep Convolutional Neural Network (DCNN) to forecast the VIX, integrating a conditional variance model in the final layer to jointly predict mean and variance. The variance is embedded within the probability-based loss

function as a way to reduce uncertainty. Compared to a standard DCNN, the model demonstrates reduced error in point predictions. Horenko, Marchenko and Gagliardini (2020) propose a multivariate nonparametric regime-switching model (TV-Entropy) based on the maximum entropy principle, applying it to forecast stock indices and estimate VaR. Compared to GARCH, TV-Entropy achieves superior Bayesian Information Criterion (BIC) scores, and better calibrated unconditional coverage on 95% and 99% confidence intervals. Sharma, Elvira, Chouzenoux and Majumdar (2021) introduce Recurrent Dictionary Learning (RDL), which incorporates a Kalman filter with smoothing algorithms to generate distributional forecasts for stocks. The model outperforms LSTM, CNN, and ARIMA in both point forecasts and next-day trend classification, while no explicit assessment of uncertainty is conducted. Wang, Wang, Yuan, Wang and Shi (2020) propose a Leave-One-Out Cross-Conformal Predictive System (LOO-CCPS) combined with Regularized Extreme Learning Machine (RELM) to produce cumulative distribution functions (CDFs) for different assets. The model facilitates uncertainty estimation through prediction intervals derived from quantiles. The authors validate the estimates by evaluating the frequency of which values fall within the predicted quantiles, achieving superior performance compared to benchmark systems. Park, Jung, Eom and Lee (2024) introduce a Risk-Sensitive Multiagent network (RSMAN) for uncertainty aware portfolio management. The risk-sensitive agents forecast asset returns with corresponding uncertainty, directly used for portfolio construction. Compared to portfolio strategies like equally weighting, OLMAR and minimum-variance, the RSMAN is superior in terms of Sharpe ratio, annualized return and max drawdown.

In Eğrioğlu and Fildes (2020), a feed-forward neural network is initially trained without incorporating probabilistic elements. The residuals from this model are then used to generate numerous “bootstrapped" samples by perturbing returns with values drawn from the residual distribution. A separate neural network is trained on each bootstrapped sample, enabling the ensemble of models to produce distributional forecasts. The approach aims to capture total uncertainty, as the residuals reflect both epistemic and aleatoric components. However, the method assumes a constant residual distribution, ignoring the heteroscedastic nature of financial time series. Additionally, the model fails to outperform a random walk in return prediction and lacks comparison with traditional models for evaluating uncertainty estimates.

3.2. Output

While the majority of models in the sample focus on predicting returns or prices, often incorporating uncertainty estimates, some studies use volatility or volatility proxies as target variables. Figure 7 illustrates a breakdown of all articles in the sample by model output.

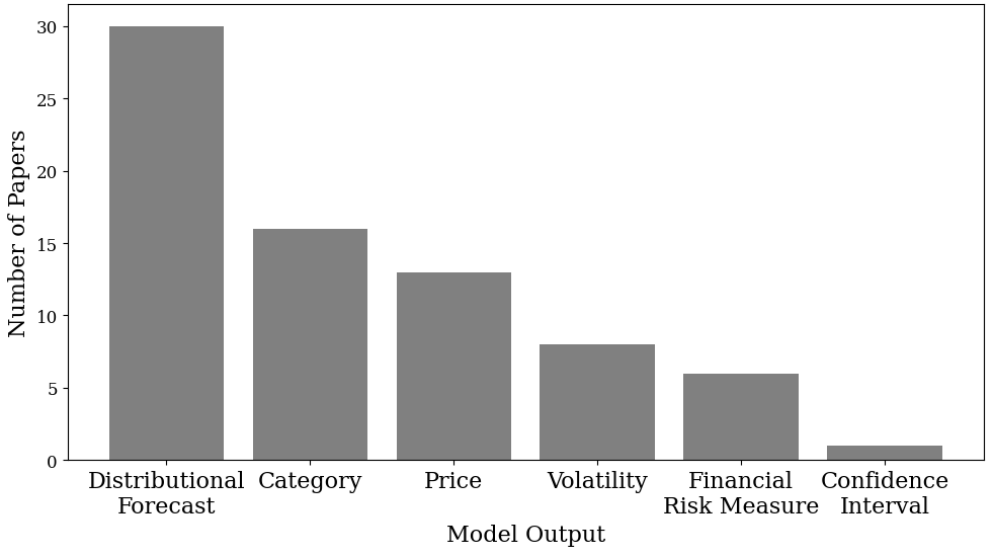

Figure 7: Distribution of model output categories.

3.2.1. Price

13 articles in the sample employ probabilistic AI models to predict asset prices or returns without exploiting the probabilistic models capability to quantify uncertainty.

Jang and Lee (2018a) use a Bayesian neural network for improved regularization and generalization, outperforming linear regression and Support Vector Regression models for Bitcoin price prediction. Jang and Lee (2018b) adopt a Bayesian approach to incorporate prior knowledge about option prices, specifically encoding in the prior that forecasted prices of deep in-the-money (ITM) or out-of-the-money (OTM) options should remain close to their previous prices. Thus, they are able to make better point predictions in situations that differ significantly from the training data, but they do not use the uncertainty estimates.

Choudhury et al. (2020) and Tang et al. (2024) use Variational Autoencoders (VAEs) as a preprocessing step for LSTM and transformer models, respectively, which in turn generates point predictions for stock prices. By first mapping the input to a lower-dimensional latent space, noise can be reduced, making the regression task easier. Choudhury et al. (2020) do not mention why a VAE was preferred over a conventional autoencoder (AE). Tang et al. (2024) suggest that VAEs perform better than AEs for noisy data, but neither their study nor the articles they cite empirically benchmark VAEs against AEs.

In Sharma et al. (2021), a clear motivation for the chosen Recurrent Dictionary Learning (RDL) structure is its ability to make predictions with uncertainty, and that a knowledgeable trader may use the uncertainty estimates to make better trading decisions. However, produced uncertainty estimates are not presented or analyzed, and there is no evaluation of whether they in fact are useful. Daniali et al. (2021) combine a CNN model with a conditional variance layer. The probabilistic output is not used, but the results show that the proposed model outperforms a traditional CNN in terms of point predictions.

In Vaiyapuri and Perumal (2016) and Li and Cheng (2010), predicted probabilities for different states are used to estimate a predicted price without any uncertainty quantification. Both articles outperform benchmarks in terms of point prediction accuracy.

The remaining five papers do not present a clear rationale for using a probabilistic AI model. Žmuk and Jošić (2020), Park et al. (2014) and Papaioannou et al. (2022) test several deterministic and Gaussian process regression (GPR) models. Both Park et al. (2014) and Papaioannou et al. (2022) show GPR as the best performing model, while in Žmuk and Jošić (2020) results are mixed. In Li et al. (2020), a VAE is used to "relief the curse of dimensions", but the authors do not explain why a *variational* autoencoder is used, rather than a traditional autoencoder. Salama (2024) employ a cGAN, only used to generate one sample at prediction time, thus not exploiting the model's ability to predict multiple future scenarios. The accuracy of the point predictions, measured in correlation between predicted returns and actual returns is as high as 0.999, raising questions about model overfitting or data leakage in the training process.

A robust conclusion is that while the use of probabilistic AI models for point prediction of asset prices may often appear arbitrary, several studies highlight their strong performance. Notably, Daniali et al. (2021) demonstrate that their probabilistic model outperforms an otherwise identical deterministic counterpart, and Jang and Lee (2018b) exploit Bayesian priors to improve generalization to unseen data.

3.2.2. Distributional Forecast.

Approximately half of the papers in the sample predict conditional distributions.

Parametric Distributions

17 articles use models that output parameters for an assumed statistical distribution, similar to GARCH models with an assumed distribution for standardized residuals. The majority of the studies use Gaussian Process Regression. GPR limits the distributional output to Gaussian forms and cannot by default handle heteroskedastic noise, restricting its utility in financial applications. Risk and Ludkovski (2018) works around this by introducing a conditional variance term in the GPR equation. The output distribution then becomes a combination of one Gaussian distribution representing epistemic uncertainty and one Gaussian distribution representing underlying data uncertainty. This enables portfolio VaR and CVaR estimation with an epistemic confidence interval. The results show that the estimates are of comparable quality to estimates achieved through computationally expensive nested Monte Carlo simulation, where the value of all portfolio assets are calculated for a range of possible economic scenarios. Another example of a GPR model effectively modeling volatility is the local volatility model proposed by Tegnér and Roberts (2021), which explicitly models and predicts the implied volatility surface inferred from option prices.

Law and Shawe-Taylor (2017) employs Bayesian SVR (B-SVR) with explicit error bars incorporating both model-driven (epistemic) uncertainty and intrinsic noise (volatility). However, intrinsic noise is assumed constant across

the time series, disregarding financial heteroskedasticity and reducing its utility as an uncertainty measure. The study also lacks validation of uncertainty estimates, though the B-SVR does outperform a traditional SVR in point prediction accuracy. B-SVR's epistemic uncertainty is not distribution-constrained because it is a sum of many distributions—one for each support vector—allowing for flexibility in the output distribution shape. In the paper, however, the practical implementation only considers the variance, removing any non-parametric characteristics.

Tian et al. (2023) analyze fitting errors to estimate uncertainty and construct prediction intervals, even though the underlying model itself is not probabilistic. These intervals account for both model uncertainty and asset volatility, but have uniform widths across the series, ignoring financial heteroskedasticity and limiting their risk analysis utility. Horenko et al. (2020) propose a simple model that is slightly freer in terms of the generated distributions where the user can choose how many moments to output—beyond just mean and variance. The model outperforms GARCH in terms of log likelihood and BIC. In Li et al. (2024a), the proposed DeepARA model outputs mean and variance, thus predicting both the expected returns and the volatility of stocks, but with an assumed distribution of returns. The usefulness of the uncertainty estimate is not assessed or benchmarked.

Non-Parametric Distributions

13 studies generate non-parametric distributions. Seven articles utilize BNNs, including both feed-forward and recurrent networks (Cocco et al., 2021; Hassan, 2024; Golnari et al., 2024; Soleymani and Paquet, 2022; Dixon, 2022; Chandra and He, 2021; Hortúa and Mora-Valencia, 2024). BNNs model the network weights as random variables. While the weight distributions often assume normality, complex interactions of hidden layers and multiple nodes allow for flexible output distributions. However, the uncertainty is primarily tied to model weights rather than data, meaning that the output primarily captures epistemic rather than aleatoric uncertainty. Standard Bayesian estimation techniques also allow for estimating aleatoric noise, but only under the assumption of constant variance, which is inadequate for financial time series prediction where volatility changes over time. Nevertheless, it is possible to construct a BNN that also quantifies heteroscedastic aleatoric uncertainty, for instance, by predicting variance alongside expected returns and training with appropriate loss functions. Such an approach, however, is only explored by Hortúa and Mora-Valencia (2024) and Soleymani and Paquet (2022). Eğrioğlu and Fildes (2020) employ bootstrapping to generate non-parametric confidence intervals, but ignore heteroscedasticity and the coverage probability of the produced confidence intervals is off by 55 percentage points.

As noted, GPR models are typically limited to parametric Gaussian output distributions and assume homoskedastic noise. Platanios and Chatzis (2014) overcome these constraints by incorporating heteroskedastic noise into the GPR framework and employing a Pitman-Yor process to integrate a potentially infinite set of GPR models, enabling the modeling of highly complex distributions. Despite these advancements, the authors do not explicitly analyze the shape of the resulting distributions. However, they demonstrate that the volatility estimates produced by their model align more closely with squared returns than those of GARCH. Their stock index modeling experiment with data from 1993 to 2003 shows a reduction to roughly one-tenth of GARCH's RMSE, while the forex experiment and the experiment on newer stock index data exhibit notable, though less extreme, improvements. Unfortunately, they do not test for significance or measure and benchmark other relevant metrics such as coverage probability.

Arian et al. (2022) employ a VAE to generate return samples for each stock in a portfolio, preserving correlations between assets. This is achieved by repeatedly sampling from the random variables in the latent space and passing these samples through the deterministic decoder part of the network. These samples can be used to construct non-parametric distributions for both individual assets and portfolio returns. From these distributions, the authors calculate VaR for three portfolios, outperforming traditional models in scoring functions, but failing Christoffersen's test for adequacy.

Fatouros et al. (2023) and Almeida et al. (2024) apply DeepAR to model asset and portfolio returns. DeepAR, inherently a multi-series model, outputs expected return and volatility for each asset, assuming a distributional form. However, the authors generate samples for portfolio returns where each sample includes simulated returns for every stock in the portfolio. This sampling process allows the construction of non-parametric distributions for the portfolio returns. Fatouros et al. (2023) show that the proposed model for forex portfolio VaR estimation passes both Christoffersen's conditional coverage test and the Dynamic Quantile (DQ) test. Additionally, it outperforms a diverse set of appropriate baseline models, such as GARCH, RiskMetrics (RM), Bidirectional Generative Adversarial Networks (BiGAN), Historical Simulation (HS) and the Monte Carlo method. The proposed model by Almeida et al. (2024) for cryptocurrency VaR and CVaR estimation is also extensively tested, but the results show that it is consistently outperformed by GARCH.

Lee and Seok (2021) and Vuletić et al. (2024) employ conditional Generative Adversarial Networks (cGAN) to forecast prices of stocks and stock indices. By inputting recent returns alongside generated noise vectors into the cGAN,

they produce samples representing diverse future scenarios, thereby forming a non-parametric distribution. Similarly, Park et al. (2024) use reinforcement learning and quantile regression to construct non-parametric distributions. However, all articles focus solely on the standard deviations of produced distributions, overlooking other potentially informative properties. Nonetheless, the meaningfulness of the uncertainty estimates is demonstrated by comparing the performance of trading strategies where the predicted standard deviations are taken into account to simpler strategies. However, informativeness of their uncertainty estimate is not assessed using e.g. Christoffersen's test, and neither is it benchmarked against traditional models such as GARCH.

Finally, Wang et al. (2020) apply a Conformal Predictive System (CPS) with a regularized extreme learning machine to produce non-parametric cumulative distribution functions (CDFs) of returns. Though not benchmarked against other models, the generated predictions appear reliable, with the observed quantiles closely matching expected frequencies.

3.2.3. Volatility

While volatility can be inferred from the probabilistic outputs of some models discussed in Section 3.2.2, eight articles in the sample explicitly predict volatility or its proxies.

Xing et al. (2019), Parker et al. (2021) and Platanios and Chatzis (2014) attempt to model latent, unobservable volatility directly, without relying on proxies. Xing et al. (2019) achieves this by using a negative ELBO loss function to train a proposed hybrid model, combining a VAE and a RNN with sentiment data. Theoretically, minimizing the negative ELBO enables the model to make optimal predictions for latent volatility. Model performance is evaluated through negative log-likelihood (NLL) and compared against traditional models like GARCH variants and other machine learning methods, showing consistent outperformance. Statistical tests are also conducted to assess wether the outperformance is significant, displaying strong evidence against other machine learning models, but weak evidence against GARCH, and no evidence against modified GARCH models. Parker et al. (2021) argue that the proposed Echo State Volatility Model (ESVM) provides better volatility estimates than GARCH. Platanios and Chatzis (2014) estimate volatility using a complex non-parametric distribution derived from GPR models, as detailed in Section 3.2.2. In these models, volatility is treated as heteroskedastic noise, similar to GARCH, but with seemingly higher accuracy in terms of RMSE against squared returns.

The remaining five articles predict various observable proxies for volatility. Tegnér and Roberts (2021) predict the "implied volatility surface"—that is, the implied volatility based on prices for options with different strike prices and different maturity dates—and show superior performance compared to a naive forecast. Although not fully clear, Jang and Lee (2018a) apparently use a volatility proxy. Daniali et al. (2021) use a CNN to predict the VIX, while Tian et al. (2023) use a RNN-based model to predict several volatility indices, outperforming a diverse set of benchmark models, including ARIMA and various neural networks. Lastly, Höcht et al. (2024) use a GPR to predict realized volatility with the purpose of pricing complex options.

In conclusion, the models proposed by Tegnér and Roberts (2021), Xing et al. (2019), and Platanios and Chatzis (2014) appear promising, demonstrating potential superiority over traditional models, though their evaluation methodology could be more exhaustive. Parker et al. (2021) also claim to outperform GARCH; however, the reported metrics raise concerns about the fairness of the comparison.

3.2.4. Tail Risk Measures

Caprioli et al. (2023) employ a distinct method using a VAE to generate synthetic correlation matrices as inputs for VaR calculation. Instead of deriving VaR from a single observed distribution, the VAE samples multiple plausible correlation structures to represent various market conditions. These correlation matrices are then used in a Monte Carlo simulation within a multi-factor Vasicek model to derive a distribution of portfolio losses to calculate VaR.

Among the studies estimating VaR, their approaches to evaluating the correctness of these predictions vary. The most straightforward method to assess prediction intervals and VaR estimates is to verify whether the frequency of violations—i.e., the occurrence of data points exceeding the predicted VaR—matches the chosen significance level. For instance, with a VaR estimate at a 5% significance level, approximately 5% of observed values should lie outside the predicted range. Over the short term, discrepancies may arise, but in the long term this should hold. This property can be statistically tested via an unconditional coverage test, commonly referred to as Kupiec's test (Kupiec, 1995). Fatouros et al. (2023) and Arian et al. (2022) both pass this test, even when GARCH models do not. Horenko et al. (2020) also reference Kupiec and report VaR violation frequencies, though the absence of p-values makes it unclear if the model passes the test. In the other three articles, no such test is conducted (Almeida et al., 2024; Risk and

Ludkovski, 2018; Caprioli et al., 2023). Fatouros et al. (2023) and Arian et al. (2022) conduct the conditional coverage test of Christoffersen (1998), which and the model in Arian et al. (2022) fails.

Although testing CVaR estimates is more complex, (Acerbi and Szekely, 2014) and (Du and Escanciano, 2017), among others, have developed statistical test for this purpose. However, neither of the two articles that estimate CVaR apply these tests; instead, they rely on scoring functions for evaluation. Risk and Ludkovski (2018) assess CVaR using RMSE against Harrell-Davis estimates as a proxy for the "ground truth," without benchmarking against traditional models. Almeida et al. (2024) use the Continuous Ranked Probability Score (CRPS) and demonstrate superiority relative to GARCH.

Additionally, to demonstrate that a VaR or CVaR model is an improvement over traditional approaches, appropriate scoring functions and benchmarking are essential. All six articles employ scoring functions, yet only four benchmark their models against traditional ones, with only Fatouros et al. (2023) and Horenko et al. (2020) demonstrating clear superiority.

3.2.5. Categorization

Sixteen articles in the sample propose classification-based models for financial forecasting, targeting variables such as the direction of price changes. Most machine learning classification models can output class probabilities, but these probabilities are often poorly calibrated, which means that the estimated probabilities do not match the true class proportions (Guo, Pleiss, Sun and Weinberger, 2017). If that is the case, the predictions lose their interpretability, and the probabilities provide little insight. Thus, we only include classification models that are based on probability theory in this review.

Some of the most commonly used models for classification in the sample include Hidden Markov Models (Sher et al., 2023; Park et al., 2011; Zhang et al., 2019; Su and Yi, 2022; Cao et al., 2019), Bayesian Neural Networks (Malagrino et al., 2018; Magris et al., 2023) and Probabilistic Neural Networks (Thawornwong and Enke, 2004; Lahmiri, 2011; Chandrasekara et al., 2019). Although most proposed classification models in the sample have the potential to produce well-calibrated probabilities, few articles assess whether the probabilities are actually well-calibrated. Even with a model based on probability theory, it is possible to end up with distorted class probabilities, e.g. if the model is misspecified or biased. An exception is Magris et al. (2023), where the probabilities are thoroughly assessed using Expected Calibration Error (ECE) and Expected Calibration Distance (ECD). Of these two, ECD is the more robust metric, because it measures the distance between the predicted probability distribution and the empirical distribution, rather than just comparing the average difference between predicted probabilities and actual outcomes grouped into intervals (called bins), as ECE does. Thus, ECD is less sensitive to localized performance spikes.

A drawback of using a classification model instead of a price level model is that the aleatoric uncertainty estimate will no longer correspond to volatility—i.e. the standard deviation of the price—which makes it less comparable to other risk models, and prevents the calculation of financial risk measures such as Value-at-Risk. However, Kim and Lee (2023) construct a risk measure based on the uncertainty of the classification predictions and show that it is useful for portfolio management.

3.3. Asset Classes

As conveyed in Figure 8, the predominant focus in the literature is on equities.

3.3.1. Equities

Trading strategies rely on informed beliefs about future price movements (Vuletić et al., 2024), and accurate range predictions are valuable for risk management (Li et al., 2024a), as improved uncertainty estimates can help investors make more informed decisions. Still, few studies explicitly state distinct motivation for constructing uncertainty estimates using probabilistic AI. Instead, the focus tends to be on price- or return predictions to develop trading strategies or inform investor decisions, with uncertainty estimates often presented as a secondary feature beneficial for risk management. Some studies examine how external factors impact the uncertainty of individual stocks across countries and sectors differently (Chandra and He, 2021; Soleymani and Paquet, 2022). Notably, Chandra and He (2021) select stocks from multiple countries to analyze the COVID-19 pandemic's effect on stock price fluctuations, highlighting the varying impact of global events on asset-level uncertainty and underscoring the need for robust uncertainty quantification.

Stock indices are most frequently predicted asset, with strong focus on American, European and Asian indices. Since indices are typically composed by multiple stocks from different sectors, they are generally less volatile than

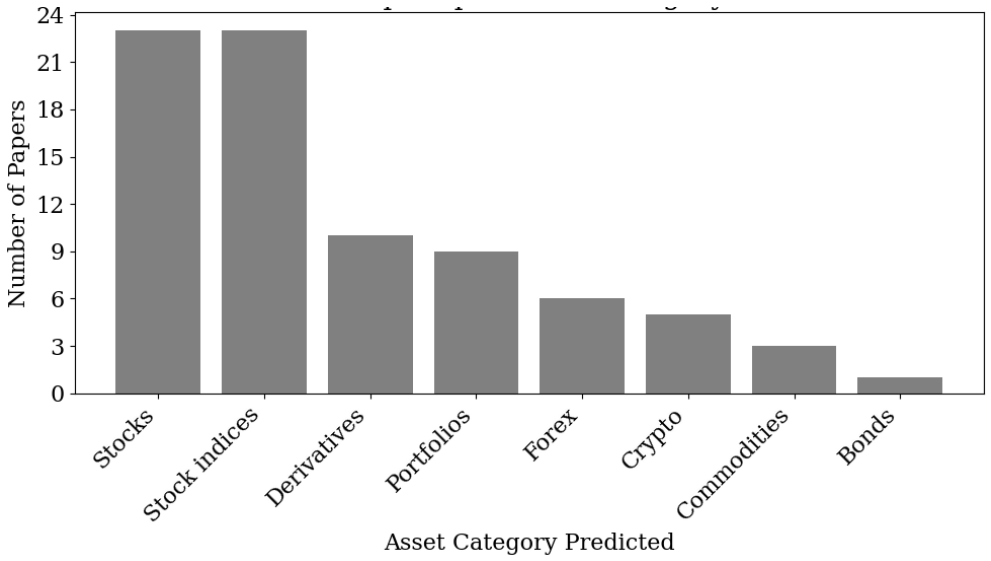

Figure 8: Asset type predicted across papers.

individual stocks and more indicative of the general state of the economy (Sezer et al., 2020). Therefore, uncertainty quantification can provide valuable insight into underlying market volatility.

Most authors primarily focus on the trading applications of accurately forecasting indices, but a subset of researchers emphasize the possibility of underlying market investigation. Suphawan et al. (2022) note that stock indices reflect the market, and reliable uncertainty estimates are therefore valuable in financial decision-making and risk management. Wang et al. (2021a) state that accurate forecasts of index fluctuation characteristics can aide government departments in timely and effectively supervising and guiding the market to avoid financial risk. In addition, with the globalization of the world economy, two studies investigate interdependencies between indices across the world (Cao et al., 2019; Malagrino et al., 2018), motivated by uncertainty estimates' importance to support risk management strategies on international scale.

3.3.2. Portfolios

Among the nine articles in the sample focusing on portfolios, where authors explicitly construct asset combinations and forecast value, returns or risk measures, the predominant motivation is to maximize portfolio returns while integrating financial risk measures to assess and manage uncertainty. As a result, articles in this category often genuinely care about risk. In addition, Risk and Ludkovski (2018) actualize regulatory compliance with Solvency II requirements within insurance for risk assessment at the 99.5% confidence-level. Kim and Lee (2023) emphasize the motivation for using probabilistic models with distributional outputs over deterministic models, as the variance of predicted distributions can signify uncertainty, enabling simultaneous maximization of returns and minimization of risk. Most studies derive quantile-based risk measures from distributional outputs, with Fatouros et al. (2023); Arian et al. (2022); Caprioli et al. (2023) focusing on VaR, and Risk and Ludkovski (2018); Li et al. (2023) also estimate CVaR.

3.3.3. Cryptocurrencies

Most researches are motivated by the largely fluctuating cryptocurrency prices and related implications for risk management. Golnari et al. (2024) note that rapid value fluctuations make accurate prediction challenging and emphasize that understanding the inherent uncertainty in predictions and price dynamics is crucial for effective risk management in investment and trading. Similarly, Almeida et al. (2024) highlight the substantial loss potential in crypto markets, underscoring the importance of understanding risk and implementing effective risk management strategies. Cocco et al. (2021) state that the high volatility of cryptocurrencies has made trading highly relevant in recent years, and suggest that speculation may be profitable.

3.3.4. Forex

Six articles in the sample forecast forex rates, usually just as one of many model applications studied (Park et al., 2011; Platanios and Chatzis, 2014; Tang et al., 2024; Li and Cheng, 2010; Papaioannou et al., 2022). As a result, explicit motivations for developing uncertainty estimates specific to forex forecasting is limited. However, Cao et al. (2019) highlight the importance of investigating cross-market influences, such as interaction between forex and stock markets, for international risk management, and uncertainty estimates as a way to understand the forex market's response to global market dynamics.

3.3.5. Derivatives

Hortúa and Mora-Valencia (2024); Daniali et al. (2021) analyze the VIX, while Tian et al. (2023) extend their analysis to include the COEVI and TYVIX. Park et al. (2014) analyze KOSPI options, Spiegeleer et al. (2018) S&P options and Tang et al. (2024) different ETF options. Law and Shawe-Taylor (2017) studies credit default swap spreads while Höcht et al. (2024) focus on capped volatility swaps.

3.3.6. Commodities

Wang and Lin (2024) forecast gold prices, and advocate the need for a probabilistic framework under extreme uncertainty. The authors argue that point estimates are insufficient in such conditions, whereas interval predictions provide meaningful insights for managing uncertainty. Law and Shawe-Taylor (2017) predict gold and crude oil prices among a range of other assets, but do not discuss underlying motivations. Li et al. (2020) forecast soybean futures, emphasizing the importance of accurate predictions and uncertainty quantification to facilitate decision making in agriculture risk management and crop insurance programs, vital for policymakers and investors. They also underscore that assumptions of independent variables and normal distributions commonly enforced by traditional models do not align with the real market conditions for commodities.

3.3.7. Bonds

Law and Shawe-Taylor (2017) predict various financial assets, including 10-year U.S.Treasuries and UK Gilts. They do not, however, disclose their motivation for using a probabilistic framework for predicting the bond yields specifically.

3.4. Type of Uncertainty

Probabilistic AI models are capable of distinguishing aleatoric from epistemic uncertainty. (Depeweg, Hernandez-Lobato, Doshi-Velez and Udluft, 2018) demonstrate this by decomposing the predictive variance as follows:

$$V(y^*|x^*, \mathcal{D}) = \underbrace{\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}\left[V(y^*|x^*, \mathbf{w})\right]}_{\text{aleatoric uncertainty}} + \underbrace{V_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{D})}\left[\mathbb{E}(y^*|x^*, \mathbf{w})\right]}_{\text{epistemic uncertainty}} \tag{1}$$

where y^* is the prediction for a new input x^*, $\mathcal{D}$ is the training dataset of observed input-output pairs $\{(x_i, y_i)\}_{i=1}^N$ and $\mathbf{w}$ denotes the model weights, treated as a random vector following the posterior distribution $p(\mathbf{w}|\mathcal{D})$. The first term - aleatoric uncertainty - represents the expected variance of the prediction given the model's parameters. It accounts for randomness within the data like unpredictable fluctuations in stock prices, sensor noise, or inherent variability in real-world systems. The second term - epistemic uncertainty - measures how much our model's predictions change across different possible model configurations (i.e., different values of the model parameters $\mathbf{w}$). It reflects how uncertain we are due to lack of knowledge; for example, if there is not enough data or the model has not learned the full pattern yet. More data can reduce this component. The posterior distribution, $p(\mathbf{w}|\mathcal{D})$, is the probabilistic distribution used to capture the uncertainty in the model parameters given the observed data. It reflects updated beliefs about the possible values of $\mathbf{w}$ after incorporating the training data, balancing prior knowledge with the evidence provided by the data. See Appendix C for further details.

As evident from Figure 9, many studies neither use nor interpret the uncertainty estimates at all. When uncertainty estimates are presented, authors usually treat them as total uncertainty, without assessing whether it arises from modeling limitations (epistemic uncertainty) or from the inherent volatility of the underlying asset (aleatoric uncertainty). This distinction is crucial for investment decisions because it is important to know whether the uncertainty is due to an unreliable model or a risky asset. Furthermore, only a minority of articles evaluate the quality and usefulness of the uncertainty estimates, and even fewer compare these estimates against traditional models. The following sections explore the various ways uncertainty quantification has been used in the sample articles, the financial relevance of each

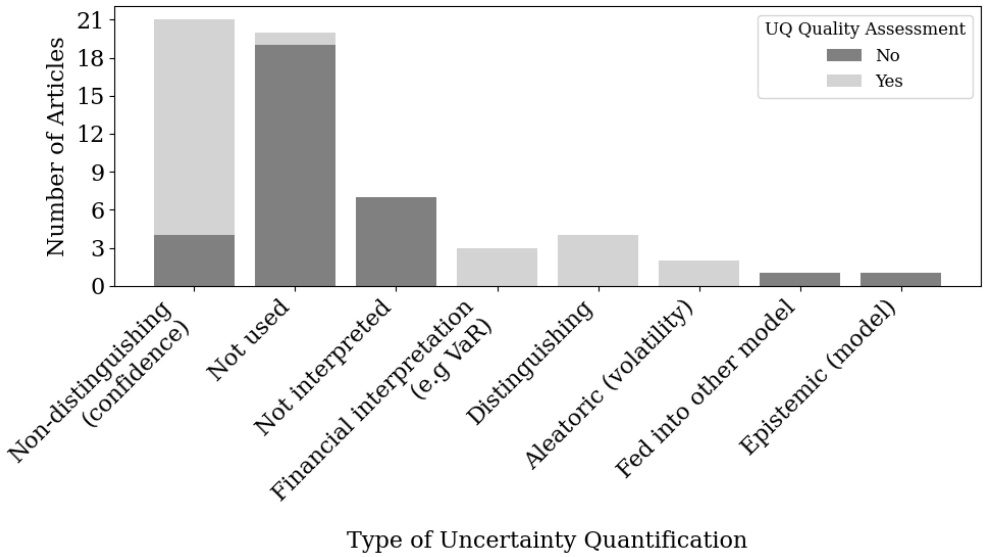

Figure 9: Authors' interpretation of the uncertainty estimates generated by the models and whether the correctness of these uncertainty estimates was assessed.

type of uncertainty, and how different uncertainty estimates have been and can be assessed. Table 3 summarizes the criteria employed for evaluating the uncertainty estimates.

3.4.1. Not Used

Among the 62 articles in the sample, 20 do not in any way mention or illustrate that their models can produce outputs with uncertainty, even though the utilized model inherently have the capability. In most cases, the authors' motivations for using probabilistic models are not clear, but some show that a probabilistic model can produce better point estimates (Daniali et al., 2021; Papaioannou et al., 2022; Park et al., 2014), while others apply probabilistic principles to improve generalization (Jang and Lee, 2018b).

3.4.2. Not Interpreted

In seven articles, the probabilistic outputs from the model are presented, but not interpreted or assigned any financial or technical meaning. Generally, this is simply because the authors do not focus on uncertainty estimation, but rather accuracy of point predictions (e.g. Spiegeleer et al. (2018)), or accuracy of category classifications (e.g. Malagrino et al. (2018); Zhang, Li and Pan (2016)).

3.4.3. Epistemic (model-uncertainty)

Epistemic uncertainty arises from uncertainty about which model to use or whether the estimated parameters are correct. Typical causes for this include a lack of enough data or omitted features (Hüllermeier and Waegeman, 2021).

Hassan (2024) is the only paper interpreting the quantified uncertainty as purely epistemic, with the proposed Bayesian LSTM model tasked with predicting Bitcoin price. Specifically, Monte Carlo dropout is used to estimate the LSTM weights—a Bayesian inference technique introduced by Gal and Ghahramani (2015). This technique involves keeping dropout active during prediction time, allowing the model to generate as many predictions as desired. These predictions form a distribution that captures the model's uncertainty about which weights are optimal. By default, this approach estimates only epistemic uncertainty, without capturing latent volatility or providing a total uncertainty estimate, unless specifically designed to do so. As with other uncertainty estimates, it is possible to test the adequacy of epistemic uncertainty estimates by constructing confidence intervals and examining the coverage, but such tests are not conducted in the article.

3.4.4. Aleatoric (volatility)

Only four articles in the sample that interprets the quantified uncertainty as exclusively aleatoric, meaning underlying data uncertainty, which in finance is referred to as latent volatility. This is of major interest within finance, as it is the primary measure for risk, and often the basis for deriving other risk measures (Brooks and Persand, 2003).

Two of the articles in the sample (Arian et al., 2022; Xing et al., 2019) do this through the use of Variational Auto-Encoders (VAE), that can generate non-parametric output distributions through repeated sampling, thereby capturing a spectrum of potential future scenarios.The two other articles use Quantile Regression Deep Learning (Wang and Lin, 2024) that quantifies aleatoric uncertainty through estimating several quantiles of the underlying distribution, and MaxEnt-modeling which is an application of the maximum entropy principle (Horenko et al., 2020).

All four papers apply relevant scoring metrics and all except Wang and Lin (2024) compare their results to GARCH and similar models. However, only Arian et al. (2022) and Horenko et al. (2020) perform statistical tests of adequacy.

3.4.5. Total Uncertainty (non-distinguishing)

Out of the 35 studies that interpret uncertainty, the majority (21) treat this as total uncertainty. In this context, the uncertainty is supposed to indicate how likely it is that the prediction is accurate. However, authors fail to identify the sources of uncertainty—epistemic and aleatoric—and it remains unclear how much uncertainty comes from each source.

Additionally, many authors use uncertainty quantification techniques that are unsuitable for determining total uncertainty in financial time series prediction. For example, six articles in this category employ probabilistic AI models based on Bayesian methods (Law and Shawe-Taylor, 2017; Cocco et al., 2021; Golnari et al., 2024; Dixon, 2022; Chandra and He, 2021; Magris et al., 2023). While Bayesian models can be used to estimate both epistemic and aleatoric uncertainty, they typically require making assumptions about the aleatoric noise, such as assuming it is normally distributed with a constant variance. As a result, these models are not ideal for capturing uncertainty in financial time series, where latent volatility changes over time. To address this limitation, models must be explicitly designed to account for such dynamics—for instance, by constructing a Bayesian neural network that predicts both an expected price and a variance. None of the six mentioned articles takes such an approach. Among the 21 articles in this category, only Platanios and Chatzis (2014) compares total uncertainty estimates to GARCH. The remaining 20 articles benchmark uncertainty estimates solely against other machine learning models or do not conduct benchmarking at all. This makes it difficult to assess whether the proposed models provide valuable contributions to the field in terms of effective uncertainty quantification.

3.4.6. Both Epistemic and Aleatoric Uncertainty

Only five articles estimate both epistemic and aleatoric uncertainty and distinguish between the two. This is beneficial in a financial context as it provides information both on the model's confidence and on the underlying riskiness of the asset. Additionally, it is beneficial in a model training context, as the researchers can focus on minimizing the epistemic uncertainty, while making the aleatoric uncertainty estimate as accurate as possible.

Risk and Ludkovski (2018) use a GPR, which naturally quantifies epistemic uncertainty, but modifies the regression equation to introduce a conditional variance term, similar to in GARCH, thus also capturing aleatoric uncertainty. This way, they can capture both sources of uncertainty while quantifying them separately. Unfortunately, they do not perform statistical adequacy tests, nor do they benchmark against GARCH, making it difficult to judge whether this approach is promising. In Hortúa and Mora-Valencia (2024), a BNN that outputs both an expected price and an aleatoric variance estimate is used for the VIX. Since the BNN naturally quantifies epistemic uncertainty, the approach enables quantification of both types of uncertainty. Calibration techniques to make the uncertainty estimates more accurate are also applied, and it is shown through calibration diagrams that the calibrated predicted quantiles are closer to the true proportion of values below each quantile compared to the uncalibrated ones. The best performing model achieves a scaling factor 0.9859, close to the optimal value of 1. Park et al. (2024) present RSMAN, a model that quantifies both aleatoric and epistemic uncertainty separately. Aleatoric uncertainty is estimated through quantile regression, while epistemic uncertainty is estimated by calculating the "distance" between the input data and the training data - representing how different the scenario they attempt to predict is from known historical scenarios. These separate estimates allow researchers to reduce epistemic uncertainty, for instance by gathering more data, while ensuring that aleatoric estimates accurately reflect inherent market volatility. The usefulness of this approach is illustrated by employing a portfolio construction strategy that takes both types of uncertainty into account, outperforming benchmark strategies. However, no other tests are conducted. Tegnér and Roberts (2021) employ GPR to predict the implied

volatility surface, which serves as an estimate of aleatoric uncertainty. Given the nature of GPR, the model also quantifies epistemic uncertainty in its predictions. However, the accuracy evaluation primarily compares predicted values to actual future implied volatilities, rather than directly assessing the uncertainty estimates. In Parker et al. (2021) the target variable is the log of squared returns, a proxy for true volatility, and the Bayesian model structure allows for capturing epistemic uncertainty. A perfect coverage probability is reported, but the 100% seemingly uncalibrated coverage of a GARCH benchmark, raise questions about the quality of the test.

In conclusion, several interesting approaches are taken to separate epistemic and aleatoric uncertainty, but the evaluation of the uncertainty estimates is limited.

3.4.7. Fed Into Other Model

Soleymani and Paquet (2022) use the probabilistic output from a BNN to model the underlying distribution of data, predicting drift and volatility parameters for each point used as input in a Feynman-Dirac integral. The authors do not assess these parameters, and do not try to model them using benchmark methods.

Table 3
Assessment Criteria for Uncertainty Estimates.

Evaluation Criteria	Description	Measures	Paper Count
Coverage Probability	Measures how often values fall within a given predictive interval	Coverage probability, Prediction interval coverage probability (PICP), Mean Coverage (MC)	11
Correlation & error metrics	Measures the relationship between predicted uncertainty and actual errors	Correlation between uncertainty and prediction error, Success rate, Negative log-likelihood (NLL), Area Under the ROC Curve (AUROC), RMSE (against squared returns), Probabilistic Trend Prediction Precision (PTPP), Quadratic Loss (QL)	8
Calibration metrics	Evaluate how well predicted probabilities or uncertainty estimates align with observed outcomes	Dynamic Quantile (DQ), Quantile loss (QL), Expected Calibration Distance (ECD), Expected Calibration Error (ECE), Root Mean Squared Calibration Error (RMSCE), Calibration diagram	7
Width-Coverage scores	Combines trade-off between interval width and coverage	Continuous Ranked Probability Score (CRPS), Average Interval Score (AIS), Winkler Score, Coverage Widthbased Criterion (CWC), Mean width divided by coverage probability	6
Interval Width metrics	Measures the width of predicted intervals	Forecasting Interval Normalized Average Width (FINAW), Prediction Interval Normalized Average Width (PINAW), Semi-interval metric, MWP (Mean width percentage)	6
Portfolio & performance metrics	Evaluates the impact of predictive uncertainty on portfolio construction and performance	SVaR, Sharpe ratio, Portfolio construction and evaluation	5
Entropy & Variance metrics	Analyzes the distribution or intervals entropy and variance	Entropy of probability distribution, Kriging variance, Mean Squared Error of Variance (MSEV), Simulation variance	3
Christoffersen's test	Evaluates the conditional coverage of predictive intervals	All: Unconditional Coverage test, Independence test, Conditional Coverage test	2
Kupiec's test	Evaluates the unconditional coverage of predictive intervals	Kupiec's test	2
Other Tests	Tests that do not fit in the aforementioned categories	Largest eigenvalue in correlation matrix test	1

4. Discussion and Implications for Further Research

From Section 3 we infer that while the field of probabilistic AI for financial time series remains relatively small and fragmented, it is in rapid development. The primary focus of existing research seems to be improving point prediction accuracy for different assets, and the potential for uncertainty estimation is underutilized. Out of the 62 articles in the sample, 35 use and interpret the uncertainty estimates generated by their models. Of those that construct uncertainty estimates, 30 somehow assess the quality of their estimates, but comprehensive assessment of model accuracy and adequacy is rare. Section 3.4 reveals that the majority of studies that do utilize uncertainty estimates (21 of 35) interpret them as total uncertainty in predictions, and do hence do not distinguish between aleatoric and epistemic uncertainty. Only nine articles explicitly interpret uncertainty estimates as volatility, five of which distinguish between aleatoric and

epistemic uncertainty while four estimate aleatoric only. No specific models or approaches seem to dominate among these. Utilization of uncertainty estimates from probabilistic AI models as a measure of volatility is thus limited. Six articles do however construct financial risk measures, such as VaR, using a probabilistic model.

A primary motivation for authors using probabilistic models across asset classes is to estimate uncertainty without having to assume any underlying distribution of the data. In spite of several studies highlighting the self-learning and noise-tolerant capabilities of probabilistic and machine learning models, we do not find clear and unambiguous motivation for incorporating uncertainty estimation with predictions. Independent of asset class, researchers center their motivation around improving risk management for investors and enhancing trading strategies through informed decision-making, but with little discussion as to how useful and appropriate the uncertainty estimates for decision making. While some studies mention asset-specific factors—such as uncertainty quantification in stock indices being tied to understanding market volatility and systematic risk, or the necessity of uncertainty estimates due to the high fluctuation characteristics of the cryptocurrency market—these are exceptions from the general lack of motivation. In many cases, authors lack a strong financial rationale for their application. An exception is the portfolio category, where researchers motivate uncertainty estimation by portfolio optimization and regulatory requirements.

Six papers estimate tail risk, with VaR being the most prevalent measure. Several studies argue that traditional parametric methods for VaR estimation are limited by explicit return distribution and linear dependency assumptions, which do not necessarily hold for financial time series (Arian et al., 2022; Fatouros et al., 2023). Probabilistic models like Variational Autoencoders (VAE) and DeepAR, utilized in several articles for VaR estimation, directly address this limitation by generating probabilistic forecasts of the entire return distribution without explicit parametric assumptions. The distributional output of the models can be used directly for tail risk estimation. Probabilistic models are thus potentially well-suited for constructing financial risk measures such as VaR or ES, and can be used to effectively address limitations of traditional models. The models are not shown to consistently outperform traditional models in the sample, largely due to failing or lacking adequacy tests, but the inherent modeling capabilities—such as arbitrary distribution shapes—hold promise for future research to create better VaR estimates.

In 16 articles in the sample, the models predict probabilities for different classes, such as price increase or decrease, rather than attempting to predict the precise price. While the magnitude of predicted changes often is lost when taking this approach, it can be highly useful for making investment decisions, and it provides an intuitive interpretation of risk. It is also possible to differentiate between epistemic and aleatoric uncertainty in classification models—particularly if the model is Bayesian or otherwise capable of representing parameter uncertainty. Under such a framework, the model might provide distributions over its parameters, allowing one to quantify how much the predicted probabilities vary as parameters are resampled from their posterior (epistemic uncertainty), and how much uncertainty remains even with fixed parameters (aleatoric uncertainty). Unfortunately, none of the articles in the sample explicitly conduct such a detailed decomposition.

Benchmarking of probabilistic model's uncertainty estimates against traditional models remains limited. Only eight papers in the sample compare uncertainty estimates to econometric models. In seven of these authors benchmark against GARCH variants, and in five cases does the proposed model clearly outperform on the reported metrics assessing uncertainty. These models are the DeepVaR proposed by Fatouros et al. (2023), the GPMCH by Platanios and Chatzis (2014), the ESVM by Parker et al. (2021) and the SAVING model by Xing et al. (2019). However, there are 12 articles in which the authors compare uncertainty estimates with another machine learning model, where only three also compare against an econometric model. All models outperform the machine learning model baselines. These findings suggest lacking and unsatisfactory benchmarking against traditional econometric models like GARCH, making it difficult to draw definitive conclusions.

When it comes to point predictions, only a minority of articles — 29 of 62 — benchmark their point predictions against traditional models, mostly ARIMA and linear regression. Of these, 23 report superior performance, demonstrating promise in accurate predictive power. However, benchmark models like ARIMA rely on correct specification, risking underperformance if misspecified, potentially inflating the apparent success of the proposed models. In contrast, only four articles benchmark their point predictions against random walk or naive forecasts, which for point prediction benchmarking effectively refer to the same method. Of these, two models fail to outperform random walk (Hortúa and Mora-Valencia, 2024; Eğrioğlu and Fildes, 2020), one demonstrates clear superiority (Papaioannou et al., 2022), while the last one marginally outperforms (Thawornwong and Enke, 2004), undermining it's reliability without significance tests. As random walk models do not require parameter estimation and thus can not be poorly specified, they serve as robust benchmarks. Additionally, they can help draw conclusions about the true predictive power of the proposed model, as a model not able to clearly outperform a random walk effectively has no predictive power. Therefore, we

suggest that if traditional econometric benchmarks are used, a random walk benchmark should also be included, to enhance evaluation rigor and mitigate risk of unreliable conclusions about predictive performance. We note that among the journal categories, as classified in Figure 4 and Appendix B, there is no clear disciplinary pattern in the use of the random walk benchmark. The few publications that include it are scattered across journals in Economics, Finance, and Business, Computer Science and Artificial Intelligence, as well as Physics and Mathematics.

A total of 46 articles benchmark their point predictions against other machine learning models, and 37 of them outperform their benchmarks. Again, this suggests a bias towards comparing proposed models to other machine learning models rather than traditional econometric models. Although summary statistics suggest that probabilistic AI models have strong point prediction capabilities, their reliability remains uncertain due to several factors: the lack of benchmarking against simple models, evidence from a few studies showing underperformance relative to a random walk, and strong indications of publication bias in the finance literature (Kim and Ji, 2015). Moreover, researchers often rely solely on accuracy metrics for model comparison, which may not adequately capture practical usefulness. For financial stakeholders, understanding the reasoning of the model might be as important as accuracy (Freeborough and van Zyl, 2022). Probabilistic AI models are in contrast to traditional models still to a large extent black boxes, making it difficult to know why they make specific predictions. In this regard, explainable AI (XAI) models designed to give understandable and interpretable explanations for their predictions is a more promising field. Integrating XAI techniques, designed to provide understandable and interpretable explanations, with probabilistic AI models is hence an interesting area for further investigation.

While probabilistic models offer appealing features for financial modeling—such as estimating non-parametric distributions and capturing non-linear patterns—their reported gains in point prediction accuracy are difficult to validate due to inadequate benchmarking practices. Some models do show promise in uncertainty estimation, but the lack of thorough comparisons, particularly against traditional models, makes it challenging to assess their true potential.

As identified in Section 3, the published research comes from journals across a range of disciplines. A relevant consideration in this respect is whether the empirical approach and results are homogeneous across them. We find that Economics, Finance, & Business journals predominantly employ Gaussian Process models, while Computer Science & Artificial Intelligence journals more often adopt technically complex Other Probabilistic Methods, with customized architectures and methodological innovations. In terms of assets, stocks and stock indices dominate across all categories, though in Computer Science & Artificial Intelligence journals individual stocks are used more than twice as often as indices, contrasting with the more balanced pattern elsewhere. Regarding model outputs, distributional forecasts dominate across all categories, consistent with the probabilistic orientation of this review. Beyond this, Economics, Finance, and Business journals uniquely complement distributional forecasts with financial risk measures, reflecting the applied importance and focus of risk management in finance. Computer Science & Artificial Intelligence journals make frequent use of classification outputs, indicative of methodological experimentation, while Engineering & Technical journals are dominated by category and price forecasts, indicating a predominant focus on applications.

This review documents that applications of probabilistic AI in financial modeling remain a relatively immature field. The novelty and rapid evolution characterizing the area also represent research opportunities within the realm of uncertainty quantification and risk estimation in finance. More specifically, we see the following dimensions as the most critical to advance the field:

Distributional forecasts and tail risk: Probabilistic AI models' capability to produce non-parametric distributions remains underutilized. Exploration of this feature with thorough testing has considerable potential and could yield more accurate predictions in volatile conditions where parametric and semi-parametric models might fall short. Probabilistic models are particularly appealing for computing tail risk, such as Value at Risk and Expected Shortfall. However, research in this area is limited and often lacks scientific rigor.

Distinguishing aleatoric and epistemic risk: Explicitly differentiating between epistemic and aleatoric uncertainty will provide clearer insights into the sources of uncertainty, distinguishing the inherent underlying asset volatility and limitations in the model. Further scientific improvements on this point is of interest both for financial decision-makers in risk-reward assessments and for regulators mandated to assess both model risk and actuarial risk in financial institutions.

Benchmarking: Complex models that aim to make accurate predictions of the conditional mean, volatility or quantiles should generally include a random walk or some other simple statistical model with low risk of misspecification as a benchmark, to clearly demonstrate that the proposed model has incremental predictive power. In general, we observe that well-established statistical time-series models often are omitted as benchmarks. This is problematic, since this class

of models, albeit having a simpler structure, is less prone to overfitting and less data-intensive than ML models. Hence, they remain relevant for practical decision-makers, and should consequently be part of scientific research evaluating advanced AI models.

Performance evaluation: The research field is currently dominated by ad-hoc evaluation methods. This includes adequacy tests, performance measures as well as statistical and economic loss functions. Furthermore, there is significant potential to improve the robustness of results across regime-switching market conditions, asset classes, geographical regions, and regulatory regimes. A general remark is that the scientific field would benefit from standardized frameworks for benchmarking and performance evaluation.

Interdisciplinary research: Few articles have authors from both finance and computer science disciplines. Arguably, this constitutes a constraint in terms of rigorous finance applications within a highly complex computer science domain. We see significant potential in combining expertise from both fields to create more robust models. In particular, hybrid models that combine probabilistic AI with traditional statistical models and physics-informed systems, remain underexplored - despite their significant potential in the context of uncertainty assessments.

Replicability: To enable improvements to existing research, authors should disclose data, code and their trained models. This will facilitate direct benchmarking of new proposed models and support scientific progress in the field.

XAI: Even if probabilistic models can output both aleatoric and epistemic uncertainty, the reasoning behind predictions and related determinants typically have limited visibility. Combining probabilistic AI with techniques from explainable AI should be a prerequisite for scientific research in this field and will improve the usefulness of probabilistic AI models as tools in practical decisions.

Lastly, we note that with the rapid advancement of generative AI, including Large Language Models (LLMs) and transformer-based architectures, the role of AI in finance is expanding beyond traditional predictive analytics. Generative AI can synthesize synthetic financial data to augment training sets for probabilistic AI models, enhancing robustness in sparse-data environments. Additionally, LLMs are increasingly used for financial document analysis, regulatory compliance automation, and sentiment-based trading strategies. While our review focuses primarily on probabilistic AI models in financial time series forecasting, future research should explore the integration of LLMs for probabilistic text-based financial decision-making, such as risk disclosures, corporate earnings analysis, and automated due diligence processes. Examining the synergy between generative AI and uncertainty-aware models presents a promising avenue for improving financial predictions and risk management.

5. Conclusion

The growing reliance on AI in financial decision-making raises critical regulatory and policy concerns, particularly regarding model transparency, fairness, and systemic risk. Regulatory bodies such as the Basel Committee on Banking Supervision and the European Commission's AI Act emphasize the need for interpretable and trustworthy AI models in finance. Probabilistic AI, with its ability to provide uncertainty estimates, aligns well with regulatory demands for risk-aware AI deployment.

In this review, we perform a systematic literature review following a SLR approach to review 62 papers on the topic of probabilistic AI in finance. We examine these papers across dimensions such as model type, output, asset class, and uncertainty type. Additionally, we provide insights into the geographical distribution of research, contributor backgrounds, and the historical development of the field. Our findings suggest that most articles on probabilistic AI claim to enhance point predictions, and few articles have an explicit focus on improving uncertainty estimation within finance. Moreover, probabilistic AI offer valuable capabilities for financial modeling, including non-parametric distribution estimation, separation of uncertainty types, and capturing non-linear dynamics. However, the lack of comprehensive benchmarking and robust evaluations, especially in comparison to traditional models, makes it difficult to assess their true performance.

An important implication of our findings is the need for more interdisciplinary collaboration. Analysis of author backgrounds indicates that research in this area is largely dominated by computer scientists, with relatively limited participation from financial experts. As a result, computer scientists often lack the domain-specific knowledge needed to effectively model financial problems, while financial researchers—despite being better positioned to address such challenges—have seldom adopted probabilistic AI techniques, likely due to technical barriers. This review serves as a starting point for bridging these divides, guiding financial researchers in adopting these methods and helping computer scientists better frame their approaches within the financial context.

Finally, our review highlights the immaturity of the field. The majority of the scientific literature has been published very recently, with few papers building on each other, and relatively few achieving high standards in modeling, testing, and interpretation. Most authors do not disclose code, which hinders reproducibility and the possibility for building on previous work. While the field is promising, it would benefit greatly from following the suggested approaches for standardizing testing, benchmarking, and interpreting estimates to fully leverage the potential of probabilistic AI for uncertainty quantification in finance.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work the authors used ChatGPT as part of the initial screening of papers. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Grammarly has been utilized to improve the language and exposition of this work.

References

Abdar, M., et al., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 76, 243–297. doi:10.1016/j.inffus.2021.05.008.

Acerbi, C., Szekely, B., 2014. Backtesting Expected Shortfall: Introducing three model-independent, non-parametric back-test methodologies for Expected Shortfall. Technical Report. MSCI Inc. URL: https://www.msci.com.

Almeida, L.M., Müller, F.M., Perlin, M.S., 2024. Risk forecasting comparisons in decentralized finance: An approach in constant product market makers. Computational Economics URL: https://doi.org/10.1007/s10614-024-10585-6, doi:10.1007/s10614-024-10585-6.

Arian, H., Moghimi, M., Tabatabaei, E., Zamani, S., 2022. Encoded value-at-risk: A machine learning approach for portfolio risk measurement. Mathematics and Computers in Simulation 202, 500–525. URL: https://doi.org/10.1016/j.matcom.2022.07.015, doi:10.1016/j.matcom.2022.07.015.

Basel Committee on Banking Supervision, 2019. Minimum capital requirements for market risk URL: https://www.bis.org/basel_framework/. revised February 2019.

Blasco, T., Sánchez, J.S., García, V., 2024. A survey on uncertainty quantification in deep learning for financial time series prediction. Neurocomputing 576, 127339. URL: https://www.sciencedirect.com/science/article/pii/S0925231224001103, doi:10.1016/j.neucom.2024.127339.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327. URL: https://www.sciencedirect.com/science/article/pii/0304407686900631, doi:10.1016/0304-4076(86)90063-1.

Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis: Forecasting and Control. Wiley Series in Probability and Statistics, John Wiley & Sons, Hoboken, New Jersey.

Brooks, C., Persand, G., 2003. Volatility forecasting for risk management. Journal of Forecasting 22, 1–22. URL: https://api.semanticscholar.org/CorpusID:154615850, doi:10.1002/for.841.

Cao, W., Zhu, W., Demazeau, Y., 2019. Multi-layer coupled hidden markov model for cross-market behavior analysis and trend forecasting. IEEE Access 7, 158563–158575. doi:10.1109/ACCESS.2019.2950437.

Caprioli, S., Cagliero, E., Crupi, R., 2023. Quantifying credit portfolio sensitivity to asset correlations with interpretable generative neural networks, in: Proceedings of the 3rd Italian Workshop on Artificial Intelligence and Applications for Business and Industries (AIABI'23), CEUR Workshop Proceedings, Rome, Italy. URL: https://ceur-ws.org/Vol-XXX/paper9.pdf, doi:10.48550/arXiv.2309.08652.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering 137, 106024. doi:10.1016/j.cie.2019.106024.

Cavacini, A., 2015. What is the best database for computer science journal articles? Scientometrics 102, 2059–2071. doi:10.1007/s11192-014-1506-1.

Chandra, R., He, Y., 2021. Bayesian neural networks for stock price forecasting before and during covid-19 pandemic. PLOS ONE 16, e0253217. doi:10.1371/journal.pone.0253217.

Chandra, R., Simmons, J., 2023. Bayesian Neural Networks via MCMC: A Python-Based Tutorial. arXiv preprint arXiv:2304.02595 doi:10.48550/arXiv.2304.02595.

Chandrasekara, V., Tilakaratne, C., Mammadov, M., 2019. An improved probabilistic neural network model for directional prediction of a stock market index. Applied Sciences 9, 5334. URL: https://www.mdpi.com/2076-3417/9/24/5334, doi:10.3390/app9245334.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. arXiv preprint arXiv:1406.1078v3 URL: https://arxiv.org/abs/1406.1078v3, doi:10.48550/arXiv.1406.1078.

Choudhury, A.R., Abrishami, S., Turek, M., Kumar, P., 2020. Enhancing profit from stock transactions using neural networks. AI Communications 33, 75–92. doi:10.3233/AIC-200629.

Christoffersen, P.F., 1998. Evaluating interval forecasts. International Economic Review 39, 841–862. URL: http://www.jstor.org/stable/2527341, doi:10.2307/2527341.

Cocco, L., Tonelli, R., Marchesi, M., 2021. Predictions of bitcoin prices through machine learning based frameworks. PeerJ Computer Science doi:10.7717/peerj-cs.413.

Daniali, S.M., Barykin, S.E., Kapustina, I.V., Khortabi, F.M., Sergeev, S.M., Kalinina, O.V., Mikhaylov, A., Veynberg, R., Zasova, L., Senjyu, T., 2021. Predicting volatility index according to technical index and economic indicators on the basis of deep learning algorithm. Sustainability 13, 14011. doi:10.3390/su132414011.

Depeweg, S., Hernandez-Lobato, J.M., Doshi-Velez, F., Udluft, S., 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning, in: Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR. Proceedings of Machine Learning Research, Stockholm, Sweden. pp. 1192–1201.

Dixon, M., 2022. Industrial forecasting with exponentially smoothed recurrent neural networks. Technometrics 64, 114–124. URL: https://doi.org/10.1080/00401706.2021.1921035, doi:10.1080/00401706.2021.1921035.

Du, Z., Escanciano, J.C., 2017. Backtesting expected shortfall: Accounting for tail risk. Management Science 63, 940–958. URL: http://dx.doi.org/10.1287/mnsc.2015.2342, doi:10.1287/mnsc.2015.2342.

Eğrioğlu, E., Fildes, R., 2020. A new bootstrapped hybrid artificial neural network approach for time series forecasting. Computational Economics 59, 1355–1383. URL: http://dx.doi.org/10.1007/s10614-020-10073-7, doi:10.1007/s10614-020-10073-7.

Fatouros, G., Makridis, G., Kotios, D., Soldatos, J., Filippakis, M., Kyriazis, D., 2023. Deepvar: A framework for portfolio risk assessment leveraging probabilistic deep neural networks. Digital Finance 5, 29–56. URL: https://doi.org/10.1007/s42521-022-00050-0, doi:10.1007/s42521-022-00050-0.

Freeborough, W., van Zyl, T., 2022. Investigating explainability methods in recurrent neural network architectures for financial time series data. Applied Sciences 12, 1427. URL: http://dx.doi.org/10.3390/app12031427, doi:10.3390/app12031427.

Gal, Y., Ghahramani, Z., 2015. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Proceedings of The 33rd International Conference on Machine Learning doi:10.48550/arXiv.1506.02142.

Gandhmal, D.P., Kumar, K., 2019. Systematic analysis and review of stock market prediction techniques. Computer Science Review 34, 100190. URL: https://www.sciencedirect.com/science/article/pii/S157401371930084X, doi:10.1016/j.cosrev.2019.08.001.

Golnari, A., Komeili, M.H., Azizi, Z., 2024. Probabilistic deep learning and transfer learning for robust cryptocurrency price prediction. Expert Systems With Applications 255, 124404. URL: https://doi.org/10.1016/j.eswa.2024.124404, doi:10.1016/j.eswa.2024.124404.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. URL: https://arxiv.org/abs/1406.2661, doi:https://doi.org/10.48550/arXiv.1406.2661, arXiv:1406.2661.

Grudniewicz, J., Slepaczuk, R., 2023. Application of machine learning in algorithmic investment strategies on global stock markets. Research in International Business and Finance 66, 102052. URL: https://doi.org/10.1016/j.ribaf.2023.102052, doi:10.1016/j.ribaf.2023.102052.

Gunnarsson, E.S., Isern, H.R., Kaloudis, A., Risstad, M., Vigdel, B., Westgaard, S., 2024. Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. International Review of Financial Analysis 93, 103221. doi:10.1016/j.irfa.2024.103221.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks URL: https://arxiv.org/abs/1706.04599, doi:10.48550/ARXIV.1706.04599.

Hassan, M.M., 2024. Bitcoin Price Prediction Using Deep Bayesian LSTM With Uncertainty Quantification: A Monte Carlo Dropout–Based Approach. Stat 13, e70001. URL: https://doi.org/10.1002/sta4.70001, doi:10.1002/sta4.70001.

Hendawy, E., McMillan, D.G., Sakr, Z.M., Shahwan, T.M., 2023. Relative informative power and stock return predictability: a new perspective from egypt. Journal of Financial Reporting and Accounting URL: http://dx.doi.org/10.1108/JFRA-02-2023-0076, doi:10.1108/jfra-02-2023-0076.

Hernández-Lobato, J.M., Adams, R.P., 2015. Probabilistic backpropagation for scalable learning of bayesian neural networks, in: Proceedings of the 32nd International Conference on Machine Learning, JMLR: Workshop and Conference Proceedings, Lille, France. pp. 1861–1869.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735, doi:10.1162/neco.1997.9.8.1735.

Horenko, I., Marchenko, G., Gagliardini, P., 2020. On a computationally-scalable sparse formulation of the multidimensional and non-stationary maximum entropy principle. arXiv preprint arXiv:2005.03253 URL: https://arxiv.org/abs/2005.03253, doi:10.48550/arXiv.2005.03253.

Hortúa, H.J., Mora-Valencia, A., 2024. Forecasting VIX using Bayesian deep learning. International Journal of Data Science and Analytics doi:10.1007/s41060-024-00562-5.

Höcht, S., Schoutens, W., Verschueren, E., 2024. On the pricing of capped volatility swaps using machine learning techniques. Quantitative Finance URL: https://doi.org/10.1080/14697688.2024.2305643, doi:10.1080/14697688.2024.2305643.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. Machine Learning 110, 457–506. URL: https://doi.org/10.1007/s10994-021-05946-3, doi:10.1007/s10994-021-05946-3.

Jang, H., Lee, J., 2018a. An empirical study on modeling and prediction of bitcoin prices with Bayesian neural networks based on blockchain information. IEEE Access 6, 5427–5437. doi:10.1109/ACCESS.2017.2779181.

Jang, H., Lee, J., 2018b. Generative Bayesian neural network model for risk-neutral pricing of American index options. Quantitative Finance 19, 587–603. doi:10.1080/14697688.2018.1490807.

Jeffrey, E., 1990. Finding structure in time. Cognitive Science 14, 179–211. URL: https://doi.org/10.1207/s15516709cog1402_1, doi:10.1207/s15516709cog1402_1.

Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M., 2022. Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users. IEEE Computational Intelligence Magazine 17, 29–47. doi:10.1109/MCI.2022.3155327.

Khattak, B.H.A., Shafi, I., Khan, A.S., Flores, E.S., Lara, R.G., Samad, M.A., Ashraf, I., 2023. A Systematic Survey of AI Models in Financial Market Forecasting for Profitability Analysis. IEEE Access , 1–18doi:10.1109/ACCESS.2023.3330156.

Kim, J., Lee, M., 2023. Portfolio optimization using predictive auxiliary classifier generative adversarial networks. Engineering Applications of Artificial Intelligence 125, 106739. URL: https://www.sciencedirect.com/science/article/pii/S0952197623009235, doi:10.1016/j.engappai.2023.106739.

Kim, J.H., Ji, P.I., 2015. Significance testing in empirical finance: A critical review and assessment. FIRN (Financial Research Network) Research Paper Series URL: https://api.semanticscholar.org/CorpusID:33194952.

Kingma, D.P., Welling, M., 2014. Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114 URL: https://arxiv.org/abs/1312.6114, doi:10.48550/arXiv.1312.6114.

Kuhrmann, M., Méndez Fernández, D., Daneva, M., 2017. On the pragmatic design of literature studies in software engineering: an experience-based guideline. Empirical Software Engineering 22, 2852–2891. doi:10.1007/s10664-016-9492-y.

Kumbure, M.M., Lohrmann, C., Luukka, P., Porras, J., 2022. Machine learning techniques and data for stock market forecasting: A literature review. Expert Systems With Applications 197, 116659. URL: https://doi.org/10.1016/j.eswa.2022.116659, doi:10.1016/j.eswa.2022.116659.

Kupiec, P.H., 1995. Techniques for verifying the accuracy of risk measurement models. The Journal of Derivatives 3, 73–84. doi:10.3905/jod.1995.407942.

Lahmiri, S., 2011. Neural networks and investor sentiment measures for stock market trend prediction. Journal of Theoretical and Applied Information Technology 27. URL: https://www.jatit.org/volumes/Vol27No1/1Vol27No1.pdf.

Law, T., Shawe-Taylor, J., 2017. Practical Bayesian support vector regression for financial time series prediction and market condition change detection. Quantitative Finance 17, 1403–1416. URL: https://doi.org/10.1080/14697688.2016.1267868, doi:10.1080/14697688.2016.1267868.

Lee, M., Seok, J., 2021. Estimation with uncertainty via conditional generative adversarial networks. Sensors 21, 6194. doi:10.3390/s21186194.

Li, A.W., Bastos, G.S., 2020. Stock market forecasting using deep learning and technical analysis: A systematic review. IEEE Access 8, 185232–185242. URL: https://doi.org/10.1109/ACCESS.2020.3030226, doi:10.1109/ACCESS.2020.3030226.

Li, H., Cui, Y., Wang, S., Liu, J., Qin, J., Yang, Y., 2020. Multivariate financial time-series prediction with certified robustness. IEEE Access 8, 109133–109143. doi:10.1109/ACCESS.2020.3001287.

Li, J., Chen, W., Zhou, Z., Yang, J., Zeng, D., 2024a. Deepar-attention probabilistic prediction for stock price series. Neural Computing and Applications 36, 15389–15406. URL: https://doi.org/10.1007/s00521-024-09916-3, doi:10.1007/s00521-024-09916-3.

Li, N., Xia, Z., Li, Y., Kuruoğlu, E.E., Jiang, Y., Xia, S.T., 2024b. Portfolio selection via graph-aware gaussian processes with generalized gaussian likelihood. IEEE Transactions on Artificial Intelligence 5, 505–515. URL: https://doi.org/10.1109/TAI.2023.3262456, doi:10.1109/TAI.2023.3262456.

Li, S.T., Cheng, Y.C., 2010. A stochastic hmm-based forecasting model for fuzzy time series. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 40, 1255–1266. doi:10.1109/TSMCB.2009.2036860.

Li, Z., Li, C., Min, L., Lin, D., 2023. Black-litterman portfolio optimization using gaussian process regression. IAENG International Journal of Applied Mathematics 53, IJAM_53_4_34. URL: https://www.researchgate.net/publication/376380457.

Lopez, J.A., 1998. Methods for evaluating value-at-risk estimates. Federal Reserve Bank of New York Economic Policy Review 4, 119–144. Available at SSRN: https://ssrn.com/abstract=1029673.

Magris, M., Shabani, M., Iosifidis, A., 2023. Bayesian bilinear neural network for predicting the mid-price dynamics in limit-order book markets. Journal of Forecasting 42, 1407–1428. doi:10.1002/for.2955.

Malagrino, L.S., Roman, N.T., Monteiro, A.M., 2018. Forecasting stock market index daily direction: A Bayesian Network approach. Expert Systems with Applications 105, 11–22. URL: https://doi.org/10.1016/j.eswa.2018.03, doi:10.1016/j.eswa.2018.03.039.

Marzi, G., Balzano, M., Caputo, A., Pellegrini, M.M., 2024. Guidelines for bibliometric-systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. International Journal of Management Reviews n/a. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijmr.12381, doi:10.1111/ijmr.12381, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijmr.12381.

Murphy, K.P., 2022. Probabilistic Machine Learning: An introduction. MIT Press.

Neal, R.M., 1995. Bayesian Learning for Neural Networks. Phd thesis. University of Toronto.

Papaioannou, P.G., Talmon, R., Kevrekidis, I.G., Siettos, C., 2022. Time-series forecasting using manifold learning, radial basis function interpolation, and geometric harmonics. Chaos 32, 083113. URL: https://doi.org/10.1063/5.0094887, doi:10.1063/5.0094887.

Park, H., Kim, N., Lee, J., 2014. Parametric models and non-parametric machine learning models for predicting option prices: Empirical comparison study over KOSPI 200 Index options. Expert Systems with Applications 41, 5227–5237. URL: https://doi.org/10.1016/j.eswa.2014.01.032, doi:10.1016/j.eswa.2014.01.032.

Park, K., Jung, H.G., Eom, T.S., Lee, S.W., 2024. Uncertainty-aware portfolio management with risk-sensitive multiagent network. IEEE Transactions on Neural Networks and Learning Systems 35, 362–374. doi:10.1109/TNNLS.2022.3174642.

Park, S.H., Lee, J.H., Lee, H.C., 2011. Trend forecasting of financial time series using pips detection and continuous hmm. Intelligent Data Analysis 15, 779–799. doi:10.3233/IDA-2011-0495.

Parker, P.A., Holan, S.H., Wills, S.A., 2021. A general bayesian model for heteroskedastic data with fully conjugate full-conditional distributions. Journal of Statistical Computation and Simulation 91, 3207–3227. URL: https://doi.org/10.1080/00949655.2021.1925279, doi:10.1080/00949655.2021.1925279.

Pascanu, R., Mikolov, T., Bengio, Y., 2013. On the difficulty of training recurrent neural networks. arXiv preprint arXiv:1211.5063v2 URL: https://arxiv.org/abs/1211.5063v2, doi:10.48550/arXiv.1211.5063.

Platanios, E.A., Chatzis, S.P., 2014. Gaussian process-mixture conditional heteroscedasticity. IEEE Transactions on Pattern Analysis and Machine Intelligence 36, 888–900. URL: https://doi.org/10.1109/TPAMI.2013.183, doi:10.1109/TPAMI.2013.183.

López de Prado, M., 2019. Beyond econometrics: A roadmap towards financial machine learning. SSRN Electronic Journal URL: http://dx.doi.org/10.2139/ssrn.3365282, doi:10.2139/ssrn.3365282.

Qin, Z., Khawar, F., Wan, T., 2016. Collective game behavior learning with probabilistic graphical models. Neurocomputing 194, 74–86. URL: http://dx.doi.org/10.1016/j.neucom.2016.01.075, doi:10.1016/j.neucom.2016.01.075.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286. URL: http://dx.doi.org/10.1109/5.18626, doi:10.1109/5.18626.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning. The MIT Press. URL: http://www.GaussianProcess.org/gpml.

Raúl, G.M., Plaza Casado, P., Prado Román, M., 2021. Market efficiency analysis using AI models based on Investors' Mood. Revista Perspectiva Empresarial 7, 10–23. URL: https://revistas.ceipa.edu.co/index.php/perspectiva-empresarial/article/view/649, doi:10.16967/23898186.649.

Rippel, M., Jánský, I., 2011. Value at risk forecasting with the ARMA-GARCH family of models in times of increased volatility. URL: https://api.semanticscholar.org/CorpusID:54988875.

Risk, J., Ludkovski, M., 2018. Sequential design and spatial modeling for portfolio tail risk measurement. SIAM Journal on Financial Mathematics 9, 1137–1174. URL: https://doi.org/10.1137/17M1158380, doi:10.1137/17M1158380.

Salama, R., 2024. Integrating spotted hyena optimization technique with generative artificial intelligence for time series forecasting. Expert Systems doi:10.1111/exsy.13681.

Salinas, D., Flunkert, V., Gasthaus, J., 2019. Deepar: Probabilistic forecasting with autoregressive recurrent networks. arXiv preprint arXiv:1704.04110v3 URL: https://arxiv.org/abs/1704.04110v3, doi:10.48550/arXiv.1704.04110.

Sezer, O.B., Gudelek, M.U., Ozbayoglu, A.M., 2020. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. Applied Soft Computing 90, 106181. doi:10.1016/j.asoc.2020.106181.

Sharma, S., Elvira, V., Chouzenoux, E., Majumdar, A., 2021. Recurrent dictionary learning for state-space models with an application in stock forecasting. Neurocomputing 450, 1–13. doi:10.1016/j.neucom.2021.03.111.

Sher, T., Rehman, A., Kim, D., Ihsan, I., 2023. Exploiting data science for measuring the performance of technology stocks. CMC: Computers, Materials & Continua 76, 2980–2994. doi:10.32604/cmc.2023.036553.

Shi, C., Zhuang, X., 2019. A study concerning soft computing approaches for stock price forecasting. Axioms 8, 116. URL: https://www.mdpi.com/2075-1680/8/4/116, doi:10.3390/axioms8040116.

Snyder, H., 2019. Literature review as a research methodology: An overview and guidelines. Journal of Business Research 104, 333–339. URL: https://www.sciencedirect.com/science/article/pii/S0148296319304564, doi:10.1016/j.jbusres.2019.07.039.

Soleymani, F., Paquet, E., 2022. Long-term financial predictions based on Feynman–Dirac path integrals, deep Bayesian networks and temporal generative adversarial networks. Machine Learning with Applications 7, 100255. doi:10.1016/j.mlwa.2022.100255.

Specht, D.F., 1990. Probabilistic neural networks. Neural Networks 3, 109–118. URL: https://www.sciencedirect.com/science/article/pii/089360809090049Q, doi:10.1016/0893-6080(90)90049-Q.

Spiegeleer, J.D., Madan, D.B., Reyners, S., Schoutens, W., 2018. Machine learning for quantitative finance: fast derivative pricing, hedging and fitting. Quantitative Finance 18, 1635–1643. URL: https://doi.org/10.1080/14697688.2018.1495335, doi:10.1080/14697688.2018.1495335.

Su, Z., Yi, B., 2022. Research on hmm-based efficient stock price prediction. Mobile Information Systems 2022, 1–8. doi:10.1155/2022/8124149.

Suhaimi, N.A.D., Abas, H., 2020. A systematic literature review on supervised machine learning algorithms. PERINTIS eJournal 10, 1–24. URL: https://perintis.org.my/ejournalperintis/index.php/PeJ/article/view/86.

Suphawan, K., Kardkasem, R., Chaisee, K., 2022. A Gaussian Process Regression Model for Forecasting Stock Exchange of Thailand. Trends in Sciences 19, 3045. URL: https://doi.org/10.48048/tis.2022.3045, doi:10.48048/tis.2022.3045.

Tang, Y., Song, Z., Zhu, Y., Yuan, H., Hou, M., Ji, J., Tang, C., Li, J., 2022. A survey on machine learning models for financial time series forecasting. Neurocomputing 512, 363–380. URL: https://api.semanticscholar.org/CorpusID:252428231, doi:10.1016/j.neucom.2022.09.003.

Tang, Z., Huang, J., Rinprasertmeechai, D., 2024. Period-aggregated transformer for learning latent seasonalities in long-horizon financial time series. PLOS ONE 19, e0308488. URL: https://doi.org/10.1371/journal.pone.0308488, doi:10.1371/journal.pone.0308488.

Tegnér, M., Roberts, S., 2021. Probabilistic machine learning for local volatility. The Journal of Computational Finance URL: http://dx.doi.org/10.21314/jcf.2021.012, doi:10.21314/jcf.2021.012.

Thawornwong, S., Enke, D., 2004. The adaptive selection of financial and economic variables for use with artificial neural networks. Neurocomputing 56, 205–232. URL: https://www.sciencedirect.com/science/article/pii/S0925231203004806, doi:10.1016/j.neucom.2003.05.001.

Tian, C., Niu, T., Wei, W., 2023. Volatility index prediction based on a hybrid deep learning system with multi-objective optimization and mode decomposition. Expert Systems with Applications 213, 119184. doi:10.1016/j.eswa.2022.119184.

Vaiyapuri, G., Perumal, T., 2016. Probabilistic and fuzzy logic based event processing for effective business intelligence. Internatianal Arab Journal of Information Technology 13, 258–266. URL: https://ccis2k.org/iajit/PDF/Vol.13,%20No.2/7137.pdf.

Vuletić, M., Prenzel, F., Cucuringu, M., 2024. Fin-gan: Forecasting and classifying financial time series via generative adversarial networks. Quantitative Finance 24, 175–199. doi:10.1080/14697688.2023.2299466.

Wang, D., Wang, P., Yuan, Y., Wang, P., Shi, J., 2020. A fast conformal predictive system with regularized extreme learning machine. Neural Networks 126, 347–361. URL: https://www.sciencedirect.com/science/article/pii/S0893608020301167, doi:10.1016/j.neunet.2020.03.022.

Wang, J., Feng, L., Li, Y., He, J., Feng, C., 2021a. Deep nonlinear ensemble framework for stock index forecasting and uncertainty analysis. Cognitive Computation 13, 1574–1592. URL: https://doi.org/10.1007/s12559-021-09961-3, doi:10.1007/s12559-021-09961-3.

Wang, J., Feng, L., Li, Y., He, J., Feng, C., 2021b. Stock index prediction and uncertainty analysis using multi-scale nonlinear ensemble paradigm of optimal feature extraction, two-stage deep learning and gaussian process regression. Applied Soft Computing 113, 107898. URL: https://doi.org/10.1016/j.asoc.2021.107898, doi:10.1016/j.asoc.2021.107898.

Wang, Y., Lin, T., 2024. A novel deterministic probabilistic forecasting framework for gold price with a new pandemic index based on quantile regression deep learning and multi-objective optimization. Mathematics 12, 29. URL: https://doi.org/10.3390/math12010029,

doi:10.3390/math12010029.

Williams Jr, R.I., Clark, L.A., Clark, W.R., Raffo, D.M., 2021. Re-examining systematic literature review in management research: Additional benefits and execution protocols. European Management Journal 39, 521–533. doi:10.1016/j.emj.2020.09.007.

Xing, F.Z., Cambria, E., Zhang, Y., 2019. Sentiment-aware volatility forecasting. Knowledge-Based Systems 176, 68–76. doi:10.1016/j.knosys.2019.03.029.

Zhang, M., Jiang, X., Fang, Z., Zeng, Y., Xu, K., 2019. High-order hidden markov model for trend prediction in financial time series. Physica A: Statistical Mechanics and its Applications 517, 1–12. doi:10.1016/j.physa.2018.10.053.

Zhang, X.d., Li, A., Pan, R., 2016. Stock trend prediction based on a new status box method and adaboost probabilistic support vector machine. Applied Soft Computing 49, 385–398. doi:10.1016/j.asoc.2016.08.026.

Žmuk, B., Jošić, H., 2020. Forecasting stock market indices using machine learning algorithms. Interdisciplinary Description of Complex Systems 18, 471–489. URL: https://doi.org/10.7906/indecs.18.4.7, doi:10.7906/indecs.18.4.7.

A. List of Abbreviations

Table A1: List of Abbreviations

Abbreviation	Definition
ADAM	Adaptive Moment Estimation
AE	Auto-Encoder
AIS	Average Internal Score
AI	Artificial Intelligence
ANN	Artificial Neural Network
ARIMA	AutoRegressive Integrated Moving Average
AUROC	Area Under the ROC Curve
BGLM	Bayesian Generalized Linear Model
B-HANN	Bootstrapped Hybrid Artificial Neural Network
BIC	Bayesian Information Criterion
BNN	Bayesian Neural Network
B-SLR	Bibliometric Systematic Literature Review
B-SVR	Bayesian Support Vector Regression
BSTS	Bayesian Structural Time-Series
B-TABL	Bayesian Temporal Augmented Bilinear Network
CDF	Cumulative Distribution Functions
cGAN	Conditional Generative Adversarial Networks
CHMM	Continuous Hidden Markov Models
CNN	Convolutional Neural Networks
CoV	Coefficient of Variation
CPS	Conformal Predictive System
CRPS	Continuous Ranked Probability Score
CVaR / ES	Conditional Value at Risk / Expected Shortfall
CWC	Coverage Widthbased Criterion
DCNN	Deep Convolutional Neural Network
DNN	Deep Neural Network
DQ	Dynamic Quantile
ECE	Expected Calibration Error
ECD	Expected Calibration Distance
EGARCH	Exponential GARCH
EMH	Efficient Market Hypothesis
ESVM	Echo State Volatility Model
EWSVM	Enhanced Weighted Support Vector Machine
FINAW	Forecasting Interval Normalized Average Width
FF-ANN	Feed-Forward Artificial Neural Network
FFNN	Feed-Forward Neural Network
GAN	Generative Adversarial Network
GA	Genetic Algorithms
GBHM	Bayesian General Heteroskedasticity Model
GMV	Global Minimum Variance
GP	Gaussian Process
GPMCH	Gaussian Process Mixture Conditional Heteroscedasticity
GPR	Gaussian Processes Regression
GARCH	Generalized Autoregressive Conditional Heteroskedasticity
GRU	Gated Recurrent Units
HMM	Hidden Markov Model
IMF	Intrinsic Mode Functions
ITM	In-The-Money
IVOL	Implied Volatility
LOB	Limit Order Books
LOO-CCPS	Leave-One-Out Cross-Conformal Predictive System
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MC	Mean Coverage
MCUB	Multi-Class Undersampling-Based Bagging
MCHMM	Multi-Layer Coupled Hidden Markov Model
MCMC	Markov Chain Monte Carlo
MLP	Multilayer Perceptron
ML	Machine Learning
MNF	Minimum Norm Filter
MSEV	Mean Squared Error of Variance
MWP	Mean Width Percentage
NLL	Negative Log-Likelihood
NSE	Nash–Sutcliffe Model Efficiency Coefficient
OLMAR	Online Moving Average Reversion
OTM	Out-Of-The-Money
PAMR	Passive-Aggressive Mean Reversion
PDF	Probability Density Functions
PGM	Probabilistic Graphical Model

Continued on next page

Table A1 – continued from previous page

Abbreviation	Definition
PICP	Prediction Interval Coverage Probability
PINAW	Prediction Interval Normalized Average Width
PIP	Perceptually Important Points
PNN	Probabilistic Neural Network
PredACGAN	Predictive Auxiliary Classifier GAN
PTPP	Probabilistic Trend Prediction Precision
QL	Quantile Loss
QR	Quantile Regression
RDL	Recurrent Dictionary Learning
RELM	Regularized Extreme Learning Machine
RNN	Recurrent Neural Network
RMSCE	Root Mean Squared Calibration Error
RMSE	Root Mean Squared Error
RSMAN	Risk-Sensitive Multi Agent Network
RT	Reparametrization Trick
SAVING	Sentiment-Aware Volatility Forecasting
SGD	Stochastic Gradient Descent
SHOA	Spotted Hyena Optimization Algorithm
SLR	Systematic Literature Review
SSA	Singular Spectrum Analysis
SSE	Sum of Squared Errors
SVM	Support Vector Machines
SVR	Support Vector Regression
TARCH	Threshold GARCH
TCN	Temporal Convolutional Network
TVaR	Tail Value at Risk
UCRP	Uniform Constant Rebalanced Portfolio
VaR	Value at Risk
VAE	Variational Autoencoder
VDM	Variational Mode Decomposition
VOGN	Variational Online Gauss-Newton
XAI	Explainable AI

Table B1
Journals by Category with Corresponding Paper Counts in the Sample.

Category	Journals
Engineering and Technical	IEEE Access (3), Journal of Theoretical and Applied Information Technology (1), CMC-Computers Materials & Continua (1), Sensors (1), Mobile Information Systems (1)
Computer Science and Artificial Intelligence	Expert Systems with Applications (4), Neurocomputing (3), Applied Soft Computing (2), International Journal of Data Science and Analytics (1), IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics (1), Intelligent Data Analysis (1), AI Communications (1), The International Arab Journal of Information Technology (1), IEEE Transactions on Pattern Analysis and Machine Intelligence (1), Cognitive Computation (1), Expert Systems (1), IEEE Transactions on Artificial Intelligence (1), Machine Learning with Applications (1), PeerJ Computer Science (1), Neural Computing and Applications (1), IEEE Transactions on Neural Networks and Learning Systems (1), Engineering Applications of Artificial Intelligence (1), Neural Networks (1), Knowledge-Based Systems (1)
Multidisciplinary and General Science	PLoS ONE (2), Interdisciplinary Description of Complex Systems (1), Applied Sciences-Basel (1), Trends in Sciences (1), Sustainability (1), Stat (1)
Economics, Finance, and Business	Quantitative Finance (5), Computational Economics (2), Journal of Financial Reporting and Accounting (1), Digital Finance (1), Journal of Forecasting (1), Journal of Computational Finance (1), Revista Perspectiva Empresarial (1), Research in International Business and Finance (1), Journal of Risk Model Validation (1), SIAM Journal on Financial Mathematics (1)
Physics and Mathematics	Mathematics (1), Communications in Applied Mathematics and Computational Science (1), Chaos (1), Physica A-Statistical Mechanics and its Applications (1), IAENG International Journal of Applied Mathematics (1), Journal of Statistical Computation and Simulation (1), Technometrics (1), Mathematics and Computers in Simulation (1)

B. Journals per Category

C. Model Description

C.1. Bayesian Neural Networks (BNNs)

A Bayesian Neural Network (BNN) extends a traditional neural network by integrating Bayesian inference principles, allowing for the modeling of uncertainty in the network parameters (Neal, 1995). Conventional neural networks for time series define a mapping from inputs x_t to outputs y_t, which may be dependent on previous inputs and outputs, using a set of trainable weights and biases w, represented by

$$y_t = f(x_t, x_{t-1}, ..., x_{t-i}, y_{t-1}, ..., y_{t-i}; w), \tag{2}$$

where f is the composition of linear transformations and non-linear activation functions across multiple layers. BNNs extend this by providing a probabilistic implementation of a standard neural network where the weights and biases are represented as random variables with probability distributions (Chandra and Simmons, 2023), allowing the model to capture parameter uncertainty.

Initially each weight is assigned a prior distribution

$$p(w) = \Pi_i p(w_i), \tag{3}$$

where $p(w)$ represents the joint prior distribution over all weights. Combined with the likelihood of observed data $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^N$ given the weights

$$p(\mathcal{D}|w) = \prod_{n=1}^{N} p(y_n|x_{1:n}, y_{1:n-1}, w), \tag{4}$$

we can create the posterior distribution over the weights using Bayes rule (Murphy, 2022, p. 46)

$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}. \tag{5}$$

Predictions for new inputs x^* are consequently made by integrating over the posterior distribution of the weights

$$p(y^*|x^*, x_{1:n}, y_{1:n}, \mathcal{D}) = \int p(y^*|x^*, x_{1:n}, y_{1:n}, w)p(w|\mathcal{D})dw. \tag{6}$$

Modeling a posterior distribution over the weights allows uncertainty in parameters to be captured by the distribution variance, enabling predictive distributions rather than single point predictions. Thus the approach enables probabilistic forecasts, making BNNs particularly suitable for uncertainty quantification (Jospin, Laga, Boussaid, Buntine and Bennamoun, 2022). Due to intractability of the exact posterior (Hernández-Lobato and Adams, 2015), approximation methods like Monte Carlo dropout and variational inference are commonly employed.

C.2. Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a probabilistic AI model that makes no specific assumptions about the functional form of the underlying data, making it well-suited for complex regression tasks. GPR is used to perform inference over functions, defining a distribution over possible functions $f(x)$ that fit the given data (Rasmussen and Williams, 2006). Formally, a Gaussian Process (GP) is defined as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \tag{7}$$

where $m(x)$ is the mean function $\mathbb{E}(f)$, and $k(x, x')$ is the covariance function, also known as the kernel, defining how function values at point x and x' effect each other:

$$k(x, x') = Cov(f(x), f(x')) \tag{8}$$

Based on the posterior distribution, Bayesian inference (5) is applied to determine the most likely function f that fits the data, making it possible to make new predictions as new data is observed (Rasmussen and Williams, 2006). GPR provides both a predictive mean $\mathbb{E}(f_*)$ and a predictive variance $\text{Var}(f_*)$ given by:

$$\text{Var}(f(x_*)) = k(x_*, x_*)\mu\mathbf{k}_*^\top(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{k}_* \tag{9}$$

where $k(x_*, x_*)$ is the prior variance of the function at the test point x_*, $\mathbf{k}_*$ is the vector of covariances between x_* and the training inputs $\mathbf{X}$, $\mathbf{K}$ is the covariance matrix of the training inputs, σ_n^2 is the noise variance, and $\mathbf{I}$ is the identity matrix. This variance quantifies the uncertainty in the prediction at x_*, with contributions from both the prior uncertainty and the information gained from the training data. In this equation, we can consider the term σ_n^2 to be the aleatoric variance, i.e. the irreducible part of the variance stemming from data uncertainty, while the rest represents epistemic variance. A significant drawback of GPR from a financial perspective is that the aleatoric noise is assumed to be constant, ignoring the heteroscedasticity of financial data. This limitation can be remedied by extending the model to account for heteroscedasticity, which in the sample, only Risk and Ludkovski (2018); Tegnér and Roberts (2021); Platanios and Chatzis (2014) do.

C.3. Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are generative models that combine principles from deep learning and variational inference to learn probabilistic representations of data. Introduced by Kingma and Welling (2014) as Auto-Encoding Variational Bayes, the model architecture differs from traditional autoencoders by modeling the encoded "latent" space as a random vector instead of a deterministic one. Similar to a conventional autoencoder, the encoder maps input vector data x to a latent space z. However, in contrast to a conventional autoencoder, z does not consist of scalars but instead represents the parameters (mean and variance) of a probability distribution $q_\phi(z|x)$ over the latent variables, where ϕ denotes the parameters of the encoder network. The decoder reconstructs the original input data from this latent representation by decoding samples $z \sim q_\phi(z|x)$ through $p_\theta(x|z)$, aiming to model the true distribution

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz, \tag{10}$$

where the decoder is parameterized by θ. However, since the computation of the exact posterior $p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)}$ is intractable, variational inference over the latent variables is employed to approximate it with $q_\phi(z|x)$ (Kingma and Welling, 2014). As the encoder outputs a distribution over the latent variables, uncertainty can be captured in the latent representation by drawing multiple samples. These samples can then be propagated through the decoder, ultimately resulting in a distribution of reconstructed outputs.

C.4. Hidden Markov Models (HMMs)

Hidden Markov Models (HMMs) are probabilistic models used to analyze sequential data with underlying unobservable structures, extending Markov chain theory (Rabiner, 1989). In finance, HMMs can be applied to model time series where market states, such as bull or bear markets, periods of low or high volatility, or other economic regimes, are not directly observable.

The classic HMM consists of a finite set of hidden states $S = \{s_1, s_2, ...s_N\}$ and a corresponding sequence of observable outputs $O = \{o_1, o_2, ...o_T\}$. The model is defined by an initial probability distribution $\pi_i = P(q_1 = s_i)$, state transition probabilities $a_{ij} = P(q_{t+1} = s_j|q_t = s_i)$ and the emission probabilities specifying the likelihood of observations given system state $b_j(o_t) = P(o_t|q_t = s_j)$ (Rabiner, 1989). Consequently, HMMs are capable of producing distributional forecasts by utilizing the predictive probability distribution of future observations

$$P(o_{T+1}|O) = \sum_{i=1}^{N} P(o_{T+1}|q_{T+1} = s_i)P(q_{T+1} = s_i|O). \tag{11}$$

The probabilistic modeling of both hidden states and observations enables the computation of confidence intervals and prediction reliability, facilitating uncertainty estimation of predictions.

C.5. Probabilistic RNN Extensions (P-RNN)

Probabilistic extensions of Recurrent Neural Networks refer to models augmenting standard RNN implementations with stochastic components, enabling generation of probabilistic forecasts. RNNs are neural networks designed to handle sequential data by maintaining hidden states that capture information about previous inputs to shape subsequent

behavior (Jeffrey, 1990), making them suitable for financial time series analysis. In a standard RNN the hidden state h_t at time step t is updated based on the current input x_t and the previous hidden state h_{t-1}:

$$h_t = \phi(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{12}$$

where ϕ is an activation function, W_{xh} and W_{hh} are weight matrices and b_h is a bias vector.

Standard RNNs suffer from the exploding and vanishing gradient problems (Pascanu, Mikolov and Bengio, 2013), which hinder long-term dependencies and make training difficult. To address this issue, advanced architectures like Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho, van Merriënboer, Gulcehre, Bahdanau, Bougares, Schwenk and Bengio, 2014) have been introduced, incorporating gating mechanisms to control the information flow.

C.6. Probabilistic Generative Adversarial Networks

Generative Adversarial Networks (GANs) refer to a class of deep learning models introduced by Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio (2014), providing a framework for estimating generative models through an adversarial process. GANs consist of two neural networks, a generator G and a discriminator D that are trained in a two-player competitive minimax game. The generator produces synthetic data that aims to be as close to real data as possible, while the discriminator tries to distinguish between synthetic and real data samples. Both networks iteratively seek to improve.

The probabilistic capabilities of GANs comes from the generators ability to map random noise $p_z(z)$, for example Gaussian, to a probability distribution of outputs similar to the real data distribution $p_{\text{data}}(x)$, enabling it to capture the uncertainty in real-world-data. The objective can formally be formulated as:

$$\min_G \max_D \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1\mu D(G(z)))] \tag{13}$$

C.7. Probabilistic Neural Networks (PNNs)

Probabilistic Neural Networks (PNNs) are a class of feedforward neural networks with four layers leveraging statistical principles for classification tasks. First introduced by Specht (1990), a PNN does not have a specific training phase similar to regular neural networks, but a pattern layer where every input vector is measured by similarity to every sample. Subsequently, this similarity measure is used to decide what class an input most likely belongs to in the summation layer using Bayes rule (Specht, 1990). PNNs are inherently probabilistic as they utilize probability density functions to produce well-calibrated posterior probabilities for every class. The class-conditional probability for an input x is given by:

$$P(x \mid C_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \exp\left(-\frac{\|x\mu x_j\|^2}{2\sigma^2} \right) \tag{14}$$

where x_j are samples from C_i, N_i number of samples in class C_i and σ is a smoothing factor. The probabilistic structure makes it possible for dynamic probability estimations based on new data and is suitable for real-time assessment and decision-making with confidence levels for classification.

D. Descriptive Table of Key Attributes for Papers in Sample

Table D1 Descriptive Table Summarizing Key Attributes for Papers in Sample

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Almeida et al., 2024)	Crypto-currency	Ethereum portfolio	History, Technical	1 day	Distribu-tional Forecast, Financial Risk Measure	DeepAR (based on autoregressive RNN)	N/A	N/A	Financial interpretation (e.g. VaR)	Yes	Continuous Ranked Probability Score (CRPS), Elicitability score for VaR	No
(Chandrasekara et al., 2019)	Stock indices, Stocks	Australian Stock Market Index (AORD), S&P 500, Sri Lankan Stock Market Index (ASPI)	Environment, History, Technical	1 day	Category	Probabilistic Neural Network (PNN)	multi-class undersampling-based bagging (MCUB)	N/A	Not interpreted	No	N/A	No
(Daniali et al., 2021)	Volatility index	Volatility Index (VIX)	History, Technical	1 day	Price, Volatility	Deep Convolutional Neural Network (DCNN)	N/A	Conditional Variance model	Not used	No	N/A	No
(Spiegeleer et al., 2018)	Derivatives	S&P 500 American Options, S&P 500 European Options	Environment, History, Technical	Not mentioned	Distribu-tional Forecast	Gaussian Process Regression (GPR)	N/A	N/A	Not interpreted	No	N/A	No
(Dixon, 2022)	Stocks	International Business Machines (IBM)	History	1 day, 5 days	Distribu-tional Forecast	Bayesian exponential smoothed RNNs (Bayes ES-RNN) (BNN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Coverage probabillty	No
(Fatouros et al., 2023)	Portfolio	AUD/USD, EUR/USD, GBP/USD, USD/JPY	History, Technical	1 day	Distribu-tional Forecast, Financial Risk Measure	DeepAR (based on autoregressive RNN)	N/A	N/A	Financial interpretation (e.g. VaR)	Yes	Christoffersen's Test, Conditional Coverage Test, Dynamic Quantile (DQ), Firm Loss, Quadratic Loss (QL), Smooth Loss, Unconditional Coverage Test	Yes
(Golnari et al., 2024)	Crypto-currency	Bitcoin (BTC), Cardano (ADA), Ethereum (ETH), Litecoin (LTC), Polkadot (DOT), Stellar (XLM), Tron (TRX)	History, Technical	5 minutes	Distribu-tional Forecast	Probabilistic Gated Recurrent Units (P-GRU)	N/A	N/A	Non-distinguishing (confidence)	No	N/A	Yes
(Grudniewicz and Slepaczuk, 2023)	Stock indices	Budapest Stock Exhange Index (BUX), Bulgaria Stock Exchange Index (SOFIX), Deutscher Aktienindex (DAX), OMX Riga Index (OMXR), OMX Tallin Index (OMXT), OMX Vilnius Index (OMXV), Prague Stock Exchange Index (PX), S&P 500, Warsaw Stock Exchange Index (WIG20)	History, Technical	20 days	Category	Bayesian Generalized Linear Model (BGLM), Naïve Bayes (NB)	N/A	N/A	Not used	No	N/A	No
(Hassan, 2024)	Crypto-currency	Bitcoin (BTC)	History	1 day	Distribu-tional Forecast	Bayesian LSTM with Monte Carlo Drouput	N/A	N/A	Epistemic (Model)	No	N/A	No
(Hendawy et al., 2023)	Stocks	Egyptian Exchange 100 Stocks	Environment, Fundamental, History, Technical	Quarterly	Distribu-tional Forecast	Gaussian Process Regression (GPR)	N/A	N/A	Not used	No	N/A	No
(Höcht et al., 2024)	Capped volatility swaps	Apple Inc. (AAPL), JPMorgan Chase & Co (JPM), S&P 500	Environment, History, Technical	Flexible	Volatility	Gaussian Process Regression (GPR)	N/A	N/A	Not used	No	N/A	No

Continued on next page

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Horenko et al., 2020)	Stock indices	Dow Jones Industrial Average (DJI), EURO STOXX 50 (STOXX), FTSE 100, Hang Seng Index (HSI), Nikkei Stock Average (NI225), S&P 500, Swiss Market Index (SMI)	History, Technical	1 day	Distribu-tional Forecast, Financial Risk Measure	TV-Entropy	N/A	N/A	Aleatoric (Volatility), Financial interpretation (e.g. VaR)	Yes	Bayesian Information Criteria (BIC), Coverage probabillty, Kupiec's test, negative log-likelihood (NLL)	Yes
(Lahmiri, 2011)	Stock indices, Stocks	Apple Inc. (AAPL), Cisco Systems (CSCO), General Electric (GE), NYSE Composite	History, Sentiment, Technical	1 day	Category	Probabilistic Neural Network (PNN)	N/A	N/A	Not interpreted	No	N/A	No
(Law and Shawe-Taylor, 2017)	Bonds, Commodities, Credit Default Swap (CDS) Spreads, Stock indices	CME Gold Front Month Futures, FTSE 100, IBM-CDS 5YR, ICE BRENT Crude Oil Front Month Futures, S&P 500, UK Gilt 10YR Bond Yield, US Treasury 10YR Bond Yield, WMT-CDS 5YR	History	1 day	Distribu-tional Forecast	Bayesian Support Vector Regression (B-SVR)	N/A	N/A	Non-distinguishing (confidence)	Yes	Correlation between uncertainty and prediction error	No
(Li et al., 2024a)	Stocks	Chinese company stocks	History	Flexible	Distribu-tional Forecast	DeepAR with attention (DeepARA)	N/A	N/A	Non-distinguishing (confidence)	Yes	Entropy of probability distribution	No
(Li et al., 2024b)	Portfolio	MSCI World Index (MSCI), New York Stock Exchange (NYSE) stock portfolio, S&P 500, Toronto Stock Exchange (TSE) stock portfolio	History, Technical	1 day	Distribu-tional Forecast	Graph-Aware Gaussian Process (G4P)	N/A	N/A	Non-distinguishing (confidence)	Yes	Portfolio construction and evaluation	Yes
(Malagrino et al., 2018)	Stock indices	Bombay Stock Exchange (BSE 30 SENSEX), Cotation Assistée en Continu (CAC 40), Deutscher Aktienindex (DAX), Dow Jones Industrial Average (DJI), FTSE 100, Hang Seng Index (HSI), Merval (MERV), NASDAQ Composite, NYSE Composite, Nikkei Stock Average (NI225), Shanghai Composite (SSE), Stockholm General (OMXS 30)	Environment, History, Technical	1 day, 2 days, 20 days	Category	Bayesian Neural Network (BNN)	N/A	N/A	Not interpreted	No	N/A	No
(Li et al., 2023)	Portfolio, Stocks	S&P 500 stock portfolio	History	50 months	Distribu-tional Forecast	Gaussian Process Regression (GPR)	N/A	Black-Litterman	Non-distinguishing (confidence)	Yes	Portfolio construction and evaluation	No
(Papaioannou et al., 2022)	Forex	AUD/USD, CAD/USD, CHF/USD, DKK/USD, EUR/USD, GBP/USD, JPY/USD, NOK/USD, NZD/USD, SEK/USD	History	1 day	Price	Gaussian Process Regression (GPR)	N/A	N/A	Not used	No	N/A	No
(Park et al., 2014)	Options	KOSPI 200 Index options	Technical	Flexible	Price	Gaussian Process Regression (GPR)	N/A	N/A	Not used	No	N/A	No
(Park et al., 2024)	Portfolio, Stocks	KOSPI stocks, NASDAQ stocks, NYSE stocks	History, Technical	Short (millisec-onds)	Distribu-tional Forecast	Risk-sensitive multiagent network (RSMAN)	N/A	N/A	Distinguishing	Yes	Portfolio construction and evaluation	No
(Parker et al., 2021)	Stock indices	Dow Jones Industrial Average (DJI)	Environment, History, Technical	Not mentioned	Volatility	Echo State Volatility Model (ESVM)	N/A	N/A	Distinguishing	Yes	Coverage probabillty, MSEV	No

Continued on next page

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Platanios and Chatzis, 2014)	Forex, Stock indices	AUD/USD, CAD/USD, CHF/USD, Canadian TSX Composite (TSX), Cotation Assistée en Continu (CAC 40), DEM/USD, DKK/USD, Deutscher Aktienindex (DAX), FRF/USD, FTSE 100, GBP/USD, JPY/USD, Nikkei Stock Average (NI225), S&P 500	History	1 day, 1 week, 30 days	Distribu-tional Forecast, Volatility	Non-parametric Bayesian mixture of Gaussian process regression models (GPMCH)	N/A	N/A	Non-distinguishing (confidence)	Yes	RMSE (against squared returns)	No
(Raúl et al., 2021)	Stock indices	IBerian Index (IBEX)	History, Sentiment	1 day	Category	Bayesian Network (BN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Success rate	No
(Risk and Ludkovski, 2018)	Portfolio	N/A	Environment, History	1 year	Distribu-tional Forecast, Financial Risk Measure	Gaussian Process Regression (GPR)	N/A	N/A	Distinguishing, Financial interpretation (e.g. VaR)	Yes	RMSE (against Harrell-Davis as ground truth)	No
(Sharma et al., 2021)	Stocks	Chinese company stocks, India stocks, UK stocks, USA stocks	History	1 day	Category, Distribu-tional Forecast	Recurrent Dictionary Learing (RDL)	N/A	N/A	Non-distinguishing (confidence)	Yes	Log-loss	No
(Suphawan et al., 2022)	Stock indices	Stock Exchange Thailand Index (SET)	History, Technical	1 day	Distribu-tional Forecast	Gaussian Process Regression (GPR)	N/A	N/A	Non-distinguishing (confidence)	No	N/A	No
(Thawornwong and Enke, 2004)	Portfolio	S&P 500 stock portfolio	Environment, Fundamental, History, Technical	1 month	Category	Probabilistic Neural Network (PNN)	N/A	N/A	Not used	No	N/A	No
(Tian et al., 2023)	Volatility index	10-Year U.S. treasury note volatility index (TYVIX), Crude Oil ETF Volatility Index (COEVI), Volatility Index (VIX)	Environment, History, Technical	Not mentioned	Distribu-tional Forecast, Volatility	Fitting error analysis	Clockwork Recurrent Neural Network (CWRNN), Cuckoo-Search-enhanced Multi-Objective Grey Wolf Optimizer (MOGWOCS)	N/A	Non-distinguishing (confidence)	Yes	PICP (Prediction interval coverage probability), Prediction Interval Normalized Average Width (PINAW), Winkler Score, coverage widthbased criterion (CWC)	No
(Wang et al., 2021a)	Stock indices	Hang Seng Index (HSI), Nikkei Stock Average (NI225)	History	1 day	Distribu-tional Forecast	Gaussian Process Regression (GPR)	Enhanced Weighted Support Vector Machine (EWSVM), Recurrent Neural Network (RNN)	Singular Spectrum Analysis (SSA)	Non-distinguishing (confidence)	Yes	Coverage probabillty, MWP (Mean width percentage), Mean width divided by coverage probability	No
(Wang et al., 2021b)	Stock indices	Dow Jones Industrial Average (DJI), NASDAQ Composite, S&P 500	History, Technical	1 day	Distribu-tional Forecast	Gaussian Process Regression (GPR)	Auto-Encoder (AE), Long Short Term Memory Neural Network (LSTM), Recurrent Neural Network (RNN), Variational Mode Decomposition (VMD)	N/A	Non-distinguishing (confidence)	Yes	MC (Mean coverage), MWP (Mean width percentage), PICP (Prediction interval coverage probability)	No

Continued on next page

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Wang and Lin, 2024)	Commodities	Gold price	Environment, History, Sentiment	Not mentioned	Confidence Interval	Quantile Regression Bi-Directional Long Short-Term Memory (QRBiLSTM)	N/A	N/A	Aleatoric (Volatility)	Yes	PICP (Prediction interval coverage probability), Prediction Interval Normalized Average Width (PINAW), Quantile loss (QL), Semi-interval metric, average interval score (AIS)	No
(Zhang et al., 2016)	Stocks	NASDAQ stocks, Second-board Market of Shenzhen Stock Exchange (SZSE) stocks	History, Technical	Flexible	Category	Probabilistic Support Vector Machine (PSVM)	AdaBoost, Genetic Algorithm (GA)	N/A	Not interpreted	No	N/A	No
(Žmuk and Jošić, 2020)	Stock indices	Deutscher Aktienindex (DAX), Dow Jones Industrial Average (DJI), NASDAQ Composite, Nikkei Stock Average (NI225), S&P 500	History, Technical	1 month, 1 week, 2 weeks	Price	Gaussian Process Regression (GPR)	N/A	N/A	Not used	No	N/A	No
(Arian et al., 2022)	Portfolio, Stocks	Frankfurt Stock Exchange (FSE) stock portfolio, London Stock Exchange (LSE) stock portfolio, S&P 500	History, Technical	1 day	Distribu-tional Forecast, Financial Risk Measure	Variational Auto-encoder (VAE)	N/A	N/A	Aleatoric (Volatility), Financial interpretation (e.g. VaR)	Yes	Christoffersen's Test, Conditional Coverage Test, Independence Test, Kupiec's test, Lopez' loss function, Unconditional Coverage Test	Yes
(Cao et al., 2019)	Forex, Stock indices	S&P 500	History, Technical	1 week	Category	Multi-Layer Coupled Hidden Markov Model (MCHMM)	N/A	N/A	Not interpreted	No	N/A	No
(Caprioli et al., 2023)	Portfolio	Total Market Index Emerging Markets, Total Market Index Europe, Total Market Index Italy, Total Market Index US	History, Technical	1 year	Financial Risk Measure	Variational Auto-encoder (VAE)	N/A	multi-factor Vasicek model	Financial interpretation (e.g. VaR)	Yes	Te largest eigenvalue of the correlation matrix	No
(Chandra and He, 2021)	Stocks	3M Company(MMM), China Spacesat Company Limited (600118.SS), Commonwealth Bank of Australia (CBA.AX), Daimler AG (DAI. DE)	History	1 week	Distribu-tional Forecast	Langevin-gradient Bayesian neural networks (BNN) with parallel tempering Markov Chain Monte Carlo (MCMC)	N/A	N/A	Non-distinguishing (confidence)	No	Confidence interval	Yes
(Choudhury et al., 2020)	Stocks	Adobe (ADBE), Amazon (AMZN), Apple Inc. (AAPL), Cerner Corporation (CERN), Costco (COST), Facebook (FB), Fastenal Company (FAST), Google (GOOG), Hasbro Inc (HAS), IDEXX Laboratories Inc. (IDXX), Intel (INTC)	History, Technical	7 minutes	Price	Variational Auto-encoder (VAE)	Long Short Term Memory Neural Network (LSTM)	N/A	Not used	No	N/A	No
(Cocco et al., 2021)	Crypto-currency	Bitcoin (BTC), Ethereum (ETH)	Technical	1 day, 1 month, 2 weeks	Distribu-tional Forecast	BNN-SVR, Bayesian Neural Network (BNN)	N/A	N/A	Non-distinguishing (confidence)	No	N/A	No

Continued on next page

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Eğrioğlu and Fildes, 2020)	Stock indices	S&P 500	History	1 day	Distribu-tional Forecast	Bootstrapped Hybrid Artificial Neural Network (B-HANN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Reliability Evaluation (RE), lower bound closeness (LBC), mean of closeness (MC), sharpness evaluation (SE), upper bound closeness (LBC)	No
(Vaiyapuri and Perumal, 2016)	Stocks	N/A	History, Technical	Not mentioned	Price	PECEP (Combo of Complex Event Processing [CEP] and Probabilistic Fuzzy Logic [PFL])	N/A	N/A	Not used	No	N/A	No
(Hortúa and Mora-Valencia, 2024)	Derivative index	Volatility Index (VIX)	History	Not mentioned	Distribu-tional Forecast	Bayesian Neural Network (BNN), WaveNet	Temporal Convolutional Network (TCN), Transformers	N/A	Distinguishing	Yes	Calibration diagrams, PICP (Prediction interval coverage probability), RMSCE, Scaling factor	No
(Jang and Lee, 2018a)	Crypto-currency	Bitcoin (BTC)	Environment, History, Technical	Not mentioned	Price, Volatility	Bayesian neural networks (BNNs)	N/A	N/A	Not used	No	N/A	No
(Jang and Lee, 2018b)	Options	S&P 100 American put options	Technical	1 day, 1 week	Price	Generative Bayesian Neural Network (Gen-BNN)	N/A	N/A	Not used	No	N/A	No
(Kim and Lee, 2023)	Portfolio	NASDAQ 100 stocks portfolio, S&P 500 stock portfolio	History, Technical	1 month	Category	Predictive Auxiliary Classifier Generative Adversarial Networks (PredACGAN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Entropy of probability distribution	Yes
(Lee and Seok, 2021)	Stock indices	NASDAQ-100 Future Index	History, Technical	1 week	Distribu-tional Forecast	Conditional Generative Adversarial Network (CGAN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Correlation between uncertainty and prediction error, Portfolio construction and evaluation	Yes
(Li and Cheng, 2010)	Forex, Stock indices	Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX)	History, Technical	2 month	Price	Stochastic Hidden Markov Model (HMM)	N/A	N/A	Not used	No	N/A	No
(Li et al., 2020)	Commodities	Soybean Futures	Environment, Fundamental, History, Technical	Not mentioned	Price	Multimodal Variational Autoencoder (VAE)	Long Short Term Memory Neural Network (LSTM), Recurrent Neural Network (RNN)	N/A	Not used	No	N/A	No
(Magris et al., 2023)	Stocks	Nasdaq Nordic Helsinki Exchange Stocks	History, Technical	Short (milliseconds)	Category	Bayesian temporal attention augmented bilinear network (B-TABL)	N/A	N/A	Non-distinguishing (confidence)	Yes	Expected Calibration Distance (ECD), Expected Calibration Error (ECE)	No
(Park et al., 2011)	Forex, Stock indices, Stocks	Boeing (BA), EUR/JPY, EUR/USD, FTSE 100, Intel (INTC), S&P 500, US-Treasury Bill 5YR, Walmart (WMT)	History, Technical	1 day, 5 days	Category	Continuous Hidden Markov Model (CHMM)	N/A	PIPs Detection Algorithm	Not used	Yes	Probabilistic Trend Prediction Precision (PTPP)	No

Continued on next page

Reference	Asset category	Asset	Input	Horizon	Predicted	Prob. AI Model	Composed with ML Model	Composed with Trad. model	Use of UQ	UQ Quality Assessment	Assessment Criteria UQ	Code
(Qin, Khawar and Wan, 2016)	Stocks	American Stocks (e.g. AT&T Inc (T), Apple Inc. (AAPL), Boeing (BA)), Chinese Stocks (e.g CN Eastern Airlines (600115SS), CN Medicine & Health (600056SS))	Technical	1 day	Category	Probabilistic Graphical Model (PGM)	N/A	N/A	Not used	No	N/A	No
(Salama, 2024)	Stocks	Axis Bank (AXISBANK), Bharat Heavy Electricals (BHEL), Container Store Group (TSC), Maruti (MARUTI), Tata Steel (TATASTEEL), Wipro (WIPRO)	History, Technical	Short (milliseconds)	Price	Conditional Generative Adversarial Network (CGAN)	Spotted Hyena Optimization Algorithm (SHOA)	N/A	Not used	No	N/A	No
(Sher et al., 2023)	Stocks	Apple Inc. (AAPL), Broadcom Inc. (AVGO), Microsoft Corporation (MSFT), Nvidia Corporation (NVDA), Taiwan Semiconductor Manufacturing Company Limited (LSM)	History	Not mentioned	Category	Hidden Markov Model (HMM)	N/A	N/A	Not used	No	N/A	Yes
(Soleymani and Paquet, 2022)	Stocks	Advanced Micro Devices (AMD), Amazon (AMZN), Apple Inc. (AAPL), Google (GOOG), Microsoft Corporation (MSFT), Nvidia Corporation (NVDA), Pfizer (PFE), Shopify (SHOP), Walmart (WMT)	History, Technical	10 days, 15 days, 20 days, 30 days, 5 days	Distribu-tional Forecast	Deep Bayesian neural networks (BNNs)	temporal generative adversarial neural networks (t-GAN)	N/A	Fed into other model	No	N/A	No
(Su and Yi, 2022)	Stock indices	Hang Seng Index (HSI)	History, Technical	1 month, 3 month, 6 month	Category, Price	Hidden Markov Model (HMM)	k-means clustering	N/A	Not used	No	N/A	No
(Tang et al., 2024)	Forex, Options, Stock indices, Stocks	Chinese Securities Index (CSI 300), ETF-Option, Exchange rates, NASDAQ Composite, S&P 500	History, Technical	1 week, 16 days, 32 days, 64 days	Price	Variational Auto-encoder (VAE)	Attention	N/A	Not used	No	N/A	Yes
(Tegnér and Roberts, 2021)	Derivatives, Stock indices	S&P 500	Technical	1 week, 3 weeks	Distribu-tional Forecast, Volatility	Bayesian GPR	N/A	N/A	Distinguishing	Yes	N/A	Yes
(Vuletić et al., 2024)	Stock indices, Stocks	American stocks (e.g. Amazon (AMZN), Nike (NKE), Pfizer (PFE)), ETFs (e.g. XKL (IBM TER), XLB (ECL IP), XLY (AMZN HD NKE))	Environment, History, Technical	Not mentioned	Distribu-tional Forecast	Conditional Generative Adversarial Network (CGAN)	N/A	N/A	Non-distinguishing (confidence)	Yes	Portfolio construction and evaluation	No
(Wang et al., 2020)	Stocks	N/A	History, Technical	Not mentioned	Distribu-tional Forecast	Leave-One-Out Cross-Conformal Predictive System (LOO-CCPS)	N/A	N/A	Non-distinguishing (confidence)	Yes	Coverage probabillty, e continuous ranked probability score (CRPS)	No
(Xing et al., 2019)	Stocks	Aegon Ltd (AGN), Alibaba-group (BABA), Amazon (AMZN), Apple Inc. (AAPL), Goldman Sachs Group (GS), Google (GOOG), Pfizer (PFE), Stamper Oil & Gas Corp (STMP), Starbucks (SBUX), Tesla (TSLA)	History, Sentiment, Technical	1 day	Volatility	Variational Auto-encoder (VAE)	Recurrent Neural Network (RNN)	N/A	Aleatoric (Volatility)	Yes	negative log-likelihood (NLL)	No
(Zhang et al., 2019)	Stock indices	Chinese Securities Index (CSI 300), S&P 500	History, Technical	1 day	Category	Hidden Markov Model (HMM)	N/A	N/A	Not interpreted	No	N/A	No