Unlocking the Power of AI: Deep Learning of Conditional Volatility is Indispensable

Wenxuan Ma[a], Xing Yan[a,*]

[a]*Institute of Statistics and Big Data, Renmin University of China, Beijing, China*

Abstract

We demonstrate that predicting conditional volatility using deep learning is particularly effective and economically beneficial. The predicted conditional volatility and predicted risk premium, both monthly, can be used to form a double-sorted long-short portfolio cross-sectionally. It achieves an out-of-sample Sharpe ratio of approximately 3.0 under equal weights and about 1.5 under value weights, which are 1.0 and 0.4 higher, respectively, than those from single sorting with predicted risk premium alone. Additionally, we find significant and persistent negative relations between conditional volatilities and risk premiums in cross-sections, while previous research failed to reach a consensus. The negative risk-return relation helps to explain the superior performance of the aforementioned portfolios. Moreover, our investigation into the impact of firm characteristics reveals consistencies with well-known empirical findings, including momentum, short-term reversal, volatility persistence, and volatility asymmetry. The neural network structures and hyper-parameters used are specified without selection or searching, which minimizes implementation complexity and computational costs. Code and data are available, ensuring that all results are reproducible and free from data snooping biases.[1]

Keywords: Conditional Volatility Prediction, Deep Learning, Double Sorting, Long-Short Portfolios, Risk-Return Relationship, Impact of Firm Characteristics
JEL: G12, G17

1. Introduction

The recent surge in machine learning has introduced new tools for empirical asset pricing studies. Encouragingly, a growing body of research (Gu et al., 2020; Bali et al., 2020; Bianchi et al., 2021; Leippold et al., 2022; Aït-Sahalia et al., 2022; Choi et al., 2023; Cakici et al., 2023; Kelly et al., 2023) demonstrates that machine learning models can effectively predict expected excess returns of assets, or risk premiums, in academic finance terminology. This offers substantial economic benefits to investors, with long-short portfolios exhibiting impressive performance both in their original form

*Corresponding author.
Email addresses: mawenxuan@ruc.edu.cn (Wenxuan Ma), xingyan@ruc.edu.cn (Xing Yan)

[1]Our code and data for reproducing all results are provided at out GitHub page: https://github.com/xingyan-fml/DLConditionalVolatility.

and when risk-adjusted. The primary predictive signals are linked to popular factor types such as momentum, reversal, liquidity, and size (Gu et al., 2020; Leippold et al., 2022; Cakici et al., 2023). Forecasting with machine learning, especially deep learning, is inherently nonlinear and involves complex variable interactions, which traditional linear models in empirical asset pricing fail to capture.

This development opens new avenues for academic finance. Among machine learning models, deep learning—specifically, neural networks—has proven particularly effective, often yielding superior performance. However, when deep learning models successfully predict risk premiums, an important aspect is frequently overlooked: the excess return of an individual stock, as a random variable, is inadequately described by its expectation alone (the risk premium). It is essential to consider the second moment, or total volatility, which encompasses both the volatility due to risk factor exposures and idiosyncratic volatility. Since deep learning models typically lack structural interpretability, they may struggle to differentiate between these sources of volatility. Therefore, forecasting total volatility remains a valuable alternative, a task traditionally addressed through conditional heteroskedasticity models in financial econometrics.

Conditional volatility forecasting has predominantly been dominated by GARCH-type models (Engle, 1982; Bollerslev, 1986), which are linear and mainly suited for time series analysis. In this paper, we utilize nonlinear deep learning models to predict conditional volatility, with a focus on cross-sectional studies. Deep learning is well-regarded for its capability to handle high-dimensional nonlinear modeling and extract predictive information from a large set of predictor variables (LeCun et al., 2015; Goodfellow et al., 2016). We utilize a deep learning model with dual functionality: it simultaneously predicts risk premiums and conditional volatilities. This approach is justified because these quantities are coupled—the first and second moments of excess returns. When deep learning excels in forecasting risk premiums, extending it to conditional volatility forecasting becomes natural and efficient, eliminating the need for separate model frameworks for each task. To summarize, this approach offers three key contributions.

First, the economic benefits of this approach are substantial. Recent research utilizing deep learning (Gu et al., 2020; Leippold et al., 2022; Cakici et al., 2023) has shown that forecasting risk premiums alone yields significant economic gains, with high-minus-low long-short portfolios based on predicted risk premiums outperforming those from benchmark linear models. In this paper, we further enhance this by employing the double sorting method using both predicted risk premiums and predicted conditional volatilities, with conditional volatility as the controlling variable. Remarkably, the Sharpe ratios of the resulting long-short portfolios reach approximately 3.0 under the equal-weighted scheme and about 1.5 under the value-weighted scheme, which are 1.0 and 0.4 higher, respectively, than those from single sorting with risk premiums alone. We attribute this improvement to the balanced volatilities in the long positions and short positions of our portfolios, reducing overall risk. Our long-short portfolios demonstrate robust performance even during extreme market periods, such as the 2008 financial crisis and the COVID-19 crisis in 2020. Further analysis reveals a more fundamental reason: the risk premium and conditional

volatility are statistically significantly and persistently negatively correlated in cross-sections, which constitutes the second key contribution of our paper.

The second contribution of our paper is that it facilitates the investigation of the relationship between the risk premium and conditional volatility in cross-sections, a long-standing topic known as the risk-return trade-off in the literature. Previous studies have failed to reach a consensus, despite the use of varying data sets, methodologies, and focuses on either time series or cross-sections. While Merton (1973) proposed a positive relationship between risk and return theoretically, empirical findings have been inconsistent. For instance, Glosten et al. (1993); Brandt and Kang (2004); Lochstoer and Muir (2022) observed a negative relationship, whereas León et al. (2007); Ludvigson and Ng (2007); Pástor et al. (2008) reported a positive relationship. As noted by Harvey (2001), the risk-return relationship is influenced by the model specification for estimating conditional variance, resulting in varying results depending on the model applied. In our study, employing a well-established deep learning model with strong nonlinear capabilities and the ability to extract predictive information from numerous predictors, we identify significant and persistent negative relationships between the risk premium and conditional volatility in cross-sections for most months from 2001 to 2020. The reliability of these findings is partially supported by the superior performance of the long-short portfolios described earlier, which in turn helps explain their performance. Thus, these two aspects reinforce each other. Moreover, the reliability of our findings is further validated through the examination of prediction accuracy for conditional volatilities, where we demonstrate strong performance.

The third contribution of the approach in this paper is that it can improve the empirical understanding of asset returns, particularly regarding how firm characteristics influence the future risk premium and conditional volatility. Specifically, by ranking the importance of variables, we identify that the top three variable categories for predicting risk premiums are Momentum, Low Risk, and Short-Term Reversal, and the top three variable categories for predicting conditional volatilities are Size, Low Risk, and Momentum. Each of these key variables can be examined through their marginal functional relationships with the risk premium or conditional volatility. Our analysis corroborates well-established empirical findings in the literature, including the momentum effect, the short-term reversal effect, the liquidity risk effect regarding expected returns, and phenomena such as volatility persistence and asymmetric volatility. For those interested in exploring the marginal effects of any individual variables, our deep learning models allow convenient examination. Additionally, these models enable the investigation of interaction effects between predictor variables as well. In summary, our models can capture the complex relationships between a large set of predictors and target outcomes inherent in the data set.

Our empirical studies concentrate on the U.S. market, analyzing data from nearly 20,000 individual stocks over a 30-year period from 1991 to 2020, utilizing 153 firm-specific variables. We generate out-of-sample predictions for stocks' risk premiums and conditional volatilities over a 20-year period from 2001 to 2020. For the neural network structures and hyper-parameters, we specify them based on established practices from the literature and common experience, without extensive

selection or searching. This approach significantly reduces the implementation complexity and the time and resources required for computation, avoiding time-consuming hyper-parameter selection and model combination. Finally, we aim to ensure that the results are easily replicable and free from data snooping biases. The data used is sourced from Jensen et al. (2023), and the code and data acquisition method are available at https://github.com/xingyan-fml/DLConditionalVolatility.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature and delineates the connections between our work and existing research. Section 3 outlines the methodology and introduces the neural network models employed. Section 4 presents the empirical results, covering prediction accuracies, portfolio constructions based on predicted risk premiums and conditional volatilities, their relationships in cross-sections, the importance of various characteristics, and the marginal and interaction effects. Section 5 provides the concluding remarks.

2. Related Works

2.1. Machine Learning Models and Empirical Asset Pricing

In recent years, machine learning has been extensively applied to empirical asset pricing studies. Freyberger et al. (2020) utilized the adaptive group LASSO to select characteristics and estimate risk premiums in a non-parametric manner. Gu et al. (2020) employed various machine learning models, including regularized linear models, random forests, and deep neural networks, to predict risk premiums of individual stocks, demonstrating the superior performance of deep neural networks. Leippold et al. (2022) applied machine learning models to the Chinese stock market, while Cakici et al. (2023) examined expected return predictability across 46 countries using machine learning models. So far, most well-known machine learning models in textbooks have been prominently featured in the finance literature. Additionally, Rapach et al. (2010) proposed the forecast combination method to predict out-of-sample equity premiums by averaging predictions from various models, enhancing prediction accuracy and stability by smoothing the results. The forecast combination method has also been utilized in studies such as Gu et al. (2020) and Cakici et al. (2023). Besides, time-consuming hyper-parameter selection or searching is often employed.

This creates the impression that machine learning models are highly complex black-box non-parametric systems with high technical difficulty. It is a critical question to identify the most effective model for predicting stock returns and to determine whether the technical difficulty of implementation can be reduced. The studies mentioned above suggest that deep learning models are the most promising, even surpassing model combination techniques. This paper focuses exclusively on deep learning models and seeks to avoid the laborious processes of hyper-parameter selection and model combination (across different model types).

Our approach is purely empirical, focusing solely on prediction based on the data. In contrast, some other studies incorporated structures derived from financial theories into machine learning models. For example, Gu et al. (2021) introduced the autoencoder asset pricing model, a novel latent factor model that uses autoencoder neural networks to construct factors in a non-linear manner. Chen et al. (2024) combined feedforward neural networks, long short-term memory (LSTM)

networks (Hochreiter and Schmidhuber, 1997), and generative adversarial networks (Goodfellow et al., 2014) to develop a stochastic discount factor model, incorporating no-arbitrage conditions. Bryzgalova et al. (2020) created an asset pricing decision tree model with a new pruning method based on no-arbitrage constraints, while Cong et al. (2023) proposed a tree model for panel data analysis, employing iterative and global partitioning criteria instead of traditional recursive methods, also constrained by no-arbitrage conditions to generate a stochastic discount factor. At this stage, our focus remains on deep learning models without integrating these theoretical structures.

2.2. Conditional Volatility Prediction

When introducing deep learning to forecast conditional volatility (and risk premium), it is useful to contrast this with traditional models commonly used in finance literature. Conditional volatility is often forecasted using time series models, with the GARCH family being widely employed. The autoregressive conditional heteroskedasticity (ARCH) model (Engle, 1982) and its extension, the generalized autoregressive conditional heteroskedasticity (GARCH) model (Bollerslev, 1986), are renowned for capturing the linear and persistent structures of conditional volatility. Engle et al. (1987) introduced the GARCH-in-mean (GARCH-M) model, which incorporates a heteroskedasticity term into the mean equation. Nelson (1991) developed the exponential GARCH (EGARCH) model, which effectively captures asymmetries in volatility. The EGARCH model has been validated in numerous studies (Pagan and Schwert, 1990; Engle and Mustafa, 1992; Engle and Ng, 1993), often outperforming other GARCH models in volatility estimation. However, the GARCH family has notable limitations: it only incorporates historical returns as predictors, lacks flexibility due to its linear specification, and is generally less suited for cross-sectional studies. In contrast, deep learning models effectively address these limitations. They can incorporate a broader range of predictors beyond historical returns, handle complex non-linear relationships, and are well-suited for cross-sectional studies.

The GARCH-type models mentioned above rely on data of the same frequency for estimation. However, some studies incorporate mixed data of different frequencies, often including high-frequency data to estimate volatility. For instance, Ghysels et al. (2006) introduced the Mixed Data Sampling (MIDAS) approach, which uses a weighted average of lagged daily squared returns to predict monthly variance, capturing long-term volatility in financial markets. Engle et al. (2013) developed the GARCH-MIDAS model, which combines the characteristics of GARCH and MIDAS to account for both high-frequency short-term volatility and low-frequency long-term volatility. Corsi (2009) proposed the Heterogeneous Autoregressive model of Realized Volatility (HAR-RV), which integrates data of various frequencies to model long memory and fat tails. Patton and Sheppard (2015) introduced the Semivariance-HAR (SHAR) model, which decomposes daily volatilities into positive and negative semivariance. Bollerslev et al. (2016) presented the full-HARQ (HARQ-F) model, allowing all parameters to vary and incorporating estimates of measurement error variance. Bollerslev et al. (2018) proposed the Heterogeneous Exponential Realized Volatility with the Global Risk Factor (HExpGl) model, which uses exponentially weighted moving averages of high-frequency

data. While incorporating data of different frequencies significantly enhances volatility forecasting, we contend that deep learning can easily integrate various predictors of different frequencies, avoiding the need for specifying parametric model structures.

In fact, machine learning has extended its applications in volatility forecasting beyond merely predicting risk premiums. Luong and Dokuchaev (2018) integrated the HAR model with random forests to forecast realized volatility. Carr et al. (2020) utilized ridge regression, neural networks, and random forests to predict the realized variance of the S&P 500 index. Additionally, Wu and Yan (2019) employed LSTM networks to forecast conditional quantiles of financial time series, thereby predicting the first four moments of future returns. However, these studies do not fall within the scope of empirical asset pricing and do not simultaneously predict risk premiums.

2.3. Relationship between Risk Premium and Conditional Volatility

For risk-averse investors, the desirable outcome is a proportional relationship between the expected return and the total risk of an asset. However, previous studies have witnessed considerable debate regarding the relationship between the two. The majority of empirical research points to a negative correlation between conditional volatility and expected return. This issue, known as the negative risk-return trade-off puzzle, arises from inconsistencies with pricing theories. For instance, the Intertemporal Capital Asset Pricing Model (ICAPM) (Merton, 1973) posits that, assuming returns are independently and identically distributed, there should be a positive correlation between expected return and conditional variance.

Most existing studies have explored the risk-return relationship within the time series context. For instance, Glosten et al. (1993) used a modified GARCH-M model and observed a negative relationship between the conditional expected monthly return and the conditional variance of monthly return. Similarly, Brandt and Kang (2004) modeled the conditional mean and volatility of stock returns as a latent autoregressive process, uncovering a negative relationship. Lochstoer and Muir (2022) introduced a model to elucidate the weak or negative risk-return trade-off, also within the time series framework. In contrast, other studies supported a positive relationship. León et al. (2007), using MIDAS to estimate conditional variance, identified a significant positive relationship. Ludvigson and Ng (2007) employed dynamic factor analysis to estimate conditional mean and volatility, revealing a positive correlation between risk and return. Additionally, Pástor et al. (2008), using implied cost of capital (ICC), found a positive relationship between the conditional mean and variance of stock returns at both the country level and the global market level.

Harvey (2001) suggested that the risk-return relationship is sensitive to the method used for estimating conditional variance. They demonstrated this by employing nine different methods for conditional variance estimation. Consequently, the risk-return relationship may vary depending on the data period, the predictors included, and the estimation methodology. Deep learning can address these issues by handling large datasets covering an extended time period and incorporating numerous predictors. Moreover, deep learning is recognized for its effectiveness as a non-parametric, high-dimensional, and non-linear estimation method. Our analysis reveals significant and persistent negative relationships between risk premium and conditional volatility in cross-sections for

most months from 2001 to 2020. This result is further supported by the accuracy of conditional volatility predictions, and implicitly, the superior performance of double-sorted long-short portfolios constructed using both predicted risk premiums and predicted conditional volatilities. Interestingly, the negative risk-return relationship helps to explain the superior performance of these portfolios.

A similar controversy also exists in the relationship between idiosyncratic volatility and expected return in cross-sections. The majority of empirical research suggests a negative correlation between them (Ang et al., 2006, 2009; Jiang et al., 2009; Guo and Savickas, 2010). This negative correlation is referred to as an idiosyncratic volatility puzzle, as it contradicts the independence implied by pricing theories. Conversely, some studies advocate a positive correlation between idiosyncratic volatility and expected return (Merton, 1987; Fu, 2009; Bali and Cakici, 2008). This article exclusively concentrates on total volatility, omitting discussions on idiosyncratic volatility at this stage.

3. The Methodology

In this methodological section, we detail the prediction of conditional volatility using deep learning techniques. Specifically, we predict conditional volatility in conjunction with expected return or risk premium through a unified approach. While numerous studies, such as Gu et al. (2020); Leippold et al. (2022); Cakici et al. (2023), have focused on forecasting individual stock's expected returns or risk premiums with machine learning, they often overlook conditional volatility. To address this gap, we employ a deep learning method known as the Heteroscedastic Neural Network (HNN) (Nix and Weigend, 1994) and its ensemble variant, Deep Ensemble (Lakshminarayanan et al., 2017), to estimate the conditional variance of the target variable non-parametrically under the Gaussian assumption. The theoretical guarantee of this methodology is ensured by the proper scoring rule property of the loss function (Gneiting and Raftery, 2007).

In this paper, we focus exclusively on deep learning methodologies, despite the broad scope of machine learning, which includes approaches like regularized linear models, decision trees/random forests, and kernel methods. This choice is driven by the recent literature that highlights the superiority of deep learning, particularly in comparative analyses on diverse asset classes, as summarized by Kelly et al. (2023). These studies (Gu et al., 2020; Bali et al., 2020; Bianchi et al., 2021; Leippold et al., 2022; Aït-Sahalia et al., 2022; Choi et al., 2023) consistently conclude that neural networks outperform other popular machine learning techniques in terms of both forecast accuracy and economic benefits.

3.1. Model Description

Drawing inspiration from the additive prediction error model described in Gu et al. (2020), we enhance the model by revising the residual component, thus modeling the excess return of an individual stock as

$$r_{i,t+1} = \mathbb{E}_t\left[r_{i,t+1}\right] + \sqrt{\mathrm{Var}_t[r_{i,t+1}]} \cdot \varepsilon_{i,t+1}, \quad \varepsilon_{i,t+1} \sim \mathcal{N}(0,1), \tag{1}$$

Table 1: The detailed specifications for model settings and hyper-parameters of our models NN1–NN5.

	Specification
Network Structure for NN1–NN5	$[32], [32, 16], [32, 16, 8], [32, 16, 8, 4], [32, 16, 8, 4, 2]$
Activation Function at Hidden Layers	ReLU ($\max(0, x)$)
Activation Function at Output Layer	Linear for Estimating g^* and Softplus ($\log(1 + e^x)$) for h^*
Batch Size	10000
Maximum Number of Epochs	1000
Learning Rate	0.001
Other Adam Parameters	Default
L_1 Penalty Parameters	10^{-5}
Number of Models in Ensemble	10
Randomness Seeds in Ensemble	10, 1010, 2010, 3010, 4010, 5010, 6010, 7010, 8010, 9010

Notes. Additionally, the early stopping strategy is employed to prevent overfitting during the training process.

where

$$\mathbb{E}_t \left[r_{i,t+1} \right] = g^* \left(z_{i,t} \right), \tag{2}$$

and

$$\mathrm{Var}_t \left[r_{i,t+1} \right] = h^* \left(z_{i,t} \right). \tag{3}$$

This model describes the relationship between the stock excess return $r_{i,t+1}$ and the firm characteristics vector $z_{i,t}$ under the assumptions of heteroscedasticity and Gaussian noise. Stocks are indexed by $i \in I_t$ and months by $t = 1, \ldots, T$. The notation $\mathbb{E}_t[\cdot]$ represents the conditional expectation based on information up to time t, while $\mathrm{Var}_t[\cdot]$ stands for the conditional variance based on information up to time t. The goal of this method is to use deep learning to approximate the risk premium function $g^*(\cdot)$ and the conditional volatility function $\sqrt{h^*(\cdot)}$. We employ two separate neural networks to approximate g^* and h^*, yielding the estimations $\hat{g}(\cdot)$ for the risk premium $\mathbb{E}_t[r_{i,t+1}]$ and $\sqrt{\hat{h}(\cdot)}$ for the conditional volatility $\sqrt{\mathrm{Var}_t[r_{i,t+1}]}$, respectively. We assume that g^* and h^* are independent of both stock and time, utilizing the entire panel of data over a specific period to estimate them.

For the model settings and hyper-parameters of the two neural networks, we specify them based on prior experience in the literature and fix them, without engaging in model selection or hyper-parameter searching. This approach greatly demonstrates the robustness of our results, and significantly reduces implementation complexity and computational costs. We outline our specifications for model settings and hyper-parameters in Table 1. First, for the network structure, we specify five different network depths. The shallowest, NN1, has a single hidden layer with 32 neurons. NN2 has two hidden layers with 32 and 16 neurons, respectively. NN3 extends to three hidden layers with 32, 16, and 8 neurons. NN4 has four hidden layers with 32, 16, 8, and 4 neurons, while NN5 adopts the deepest structure with five hidden layers featuring 32, 16, 8, 4, and 2 neurons.

These structures are consistent with those in Gu et al. (2020). For the network estimating g^*, the ReLU activation function ($\max(0, x)$) is used at each hidden layer, with a linear function at the output layer. For the network estimating h^*, the ReLU activation is used at each hidden layer, and the Softplus activation function ($\log(1 + e^x)$) is used at the output layer to ensure the conditional variance is positive.

3.2. Learning Objective and Ensemble

For the learning objective, given the samples of $r_{i,t+1}$ and $z_{i,t}$, we train the two neural networks $g_{\theta_g}(\cdot)$ and $h_{\theta_h}(\cdot)$ with trainable parameters (coefficients) θ_g and θ_h by minimizing the negative log likelihood (NLL) loss:

$$\hat{g}, \hat{h} = \arg\min_{g_{\theta_g}, h_{\theta_h}} \frac{1}{\sum_{t=1}^{T} |I_t|} \sum_{t=1}^{T} \sum_{i \in I_t} \frac{\log h_{\theta_h}(z_{i,t})}{2} + \frac{\left(r_{i,t+1} - g_{\theta_g}(z_{i,t})\right)^2}{2h_{\theta_h}(z_{i,t})} + \text{constant}. \tag{4}$$

However, directly solving the above optimization problem using gradient descent over (θ_g, θ_h) is quite challenging. Actually, the partial derivatives of the loss with respect to θ_g or θ_h contain a scaling term of $1/h_{\theta_h}(z_{i,t})$ for every data point, causing points with low predicted $h_{\theta_h}(z_{i,t})$ to be continuously emphasized in the gradient descent process. This can lead to two potential issues: (i) gradient explosion if some $h_{\theta_h}(z_{i,t})$ values are very small, and (ii) poor estimations of $g^*(z_{i,t})$ in the regions of $z_{i,t}$ where $h_{\theta_h}(z_{i,t})$ is high.

Therefore, we shift our approach to finding a sub-optimal but reliable solution for Eqn. (4) with a two-step strategy. First, an estimation $\hat{g}$ is obtained through ordinary least squares regression. We optimize the mean squared error (MSE) loss over g_{θ_g} with L_1 regularization:

$$\hat{g} = \arg\min_{g_{\theta_g}} \frac{1}{\sum_{t=1}^{T} |I_t|} \sum_{t=1}^{T} \sum_{i \in I_t} \left(r_{i,t+1} - g_{\theta_g}(z_{i,t})\right)^2 + \lambda_g \|\theta_g\|_1, \tag{5}$$

where λ_g is a non-negative penalty parameter. In the second step, we fix $g_{\theta_g} = \hat{g}$ and find $\hat{h}$ by minimizing the NLL loss in Eqn. (4) with L_1 regularization:

$$\hat{h} = \arg\min_{h_{\theta_h}} \frac{1}{\sum_{t=1}^{T} |I_t|} \sum_{t=1}^{T} \sum_{i \in I_t} \frac{\log h_{\theta_h}(z_{i,t})}{2} + \frac{\left(r_{i,t+1} - \hat{g}(z_{i,t})\right)^2}{2h_{\theta_h}(z_{i,t})} + \lambda_h \|\theta_h\|_1, \tag{6}$$

where λ_h is the non-negative penalty parameter.

In machine learning, the regularization term is used to reduce model complexity and avoid overfitting. In addition to L_1 regularization, we also employ an early stopping strategy, which is extremely popular and useful for avoid overfitting. In short, a validation set is extracted from the training set, and the loss on this validation set is monitored. Training is stopped when this loss no longer decreases according to specific criteria during the gradient descent process. A formal description of the early stopping algorithm can be found in Gu et al. (2021). During the training of each neural network, we set the batch size to 10,000 and the maximum number of epochs to 1,000.

We use the Adam optimizer (Kingma and Ba, 2015) to solve Eqn. (5) and (6), with a learning rate of 0.001 and other parameters set to their defaults. The two penalty parameters are set to $\lambda_g = \lambda_h = 10^{-5}$. All neural network settings and hyper-parameters are listed in Table 1.

To boost the predictive performance of our model, we employ the well-known deep ensemble method (Lakshminarayanan et al., 2017) and establish ensemble out-of-sample predictions for risk premium and conditional volatility. We train $M = 10$ neural networks with the same structure and hyper-parameters, but with different randomness seeds for weight initialization. Thus, we obtain 10 different $\hat{g}$ and $\hat{h}$ estimations, denoted as $\hat{g}_j$ and $\hat{h}_j$, $j = 1, \ldots, M$. We then average these to obtain the final estimations or out-of-sample predictions of risk premium and conditional variance:

$$\bar{g}(\cdot) = \frac{1}{M}\sum_{j=1}^{M}\hat{g}_j(\cdot), \quad \bar{h}(\cdot) = \frac{1}{M}\sum_{j=1}^{M}\hat{h}_j(\cdot). \tag{7}$$

For an out-of-sample data point with firm characteristics $z_{i,t}$, we predict and denote its risk premium $\mathbb{E}_t[r_{i,t+1}]$ and conditional volatility $\sqrt{\mathrm{Var}_t[r_{i,t+1}]}$ as

$$\hat{\mu}_{i,t+1} = \bar{g}(z_{i,t}), \quad \hat{\sigma}_{i,t+1} = \sqrt{\bar{h}(z_{i,t})}, \tag{8}$$

respectively. In the following empirical studies, we analyze $\hat{\mu}_{i,t+1}$ and $\hat{\sigma}_{i,t+1}$, examining their predictability, their correlations in cross-sections, and the economic gain of predicting conditional volatility in the double-sorted long-short portfolio.

4. Empirical Studies

To the best of our knowledge, no existing works provide methodologies capable of predicting both risk premium and conditional volatility simultaneously, particularly using machine learning. When leveraging a new and powerful methodology in this paper, it is worthwhile to study the detailed implications of these predictions, as we do next. Our code and data for reproducing all the results are available at https://github.com/xingyan-fml/DLConditionalVolatility.

4.1. Data Description

We obtain monthly stock returns and stock characteristics in the U.S. market from Jensen et al. (2023)[2], covering 30 years from January 1991 to December 2020. The one-month-ahead excess return is the target variable of our deep learning models, and 153 stock characteristics are included as features, consistent with those used in Jensen et al. (2023). To ensure data quality, we exclude observations with a stock price below 1 U.S. dollar or a market capitalization below 5 million U.S. dollars. Consequently, there are a total of 18,916 stocks with 1,870,759 observations in our studies. To avoid potential errors caused by extremes, the excess return values in the training

[2]We use the SAS code at https://github.com/bkelly-lab/ReplicationCrisis to download the U.S. market data from Wharton Research Data Services (WRDS).

Table 2: The results of out-of-sample R^2_{oos} (in percentage) for neural network models with 1–5 layers (NN1–NN5).

	NN1	NN2	NN3	NN4	NN5
All	0.12	0.28	0.39	0.40	0.31
Top 30%	-0.70	0.02	0.44	0.44	0.26
Bottom 30%	0.09	0.39	0.46	0.42	0.38

Notes. In addition to the results for all data (All), we also report the results of R^2_{oos} for the top 30% by the market capitalization and for the bottom 30% by the market capitalization (sorted monthly). The training, validation, and testing are conducted using the respective part of the data.

Table 3: The results of out-of-sample average log likelihood (Gaussian) for neural network models with 1–5 layers (NN1–NN5).

	NN1	NN2	NN3	NN4	NN5
All	0.5862	0.5866	0.5887	0.5871	0.5584
Top 30%	0.8930	0.9028	0.9040	0.8992	0.8828
Bottom 30%	0.3159	0.3168	0.3192	0.3207	0.3122

Notes. In addition to the results for all data (All), we also report the results for the top 30% by the market capitalization and for the bottom 30% by the market capitalization (sorted monthly). The training, validation, and testing are conducted using the respective part of the data.

and validation data are clipped, limiting them to the range of -50% to 100% (but not in the testing data). For the stock characteristics, we replace missing values with the cross-sectional median in the corresponding month. We rank the stock characteristics in each month and map the ranks into the $[-1, 1]$ interval, as done in Kelly et al. (2019); Gu et al. (2020); Freyberger et al. (2020).

We employ a rolling-window approach for model training and out-of-sample forecasting. Initially, data from the first eleven years are divided into three sets: the training set, spanning the first seven years (1991-1997); the validation set, covering the subsequent three years (1998-2000); and the testing set, comprising the following year (2001). This partitioning method ensures that the training, validation, and testing sets do not overlap temporally. Subsequently, we roll the three sets forward by one year and repeat the model training and out-of-sample forecasting. There are a total of 20 years in the out-of-sample forecasting period, from January 2001 to December 2020.

4.2. Predictability Evaluation

We evaluate the twenty years' predictions of the risk premium, $\hat{\mu}_{i,t+1}$, and of the conditional volatility, $\hat{\sigma}_{i,t+1}$. To evaluate the predictions $\hat{\mu}_{i,t+1}$, we use the classic out-of-sample R^2 defined in Gu et al. (2020):

$$R^2_{\text{oos}} = 1 - \frac{\sum_{(i,t)\in\mathcal{T}_3} \left(r_{i,t+1} - \hat{\mu}_{i,t+1}\right)^2}{\sum_{(i,t)\in\mathcal{T}_3} r^2_{i,t+1}}, \tag{9}$$

where $\mathcal{T}_3$ indicates the testing set which does not overlap with the training set and validation set. This R^2_{oos} measures the predictive errors of all returns pooled across the firms and time, with results presented in Table 2. Meanwhile, we also assess the predictive performance of the neural network models for the top 30% large-cap stocks and for the bottom 30% small-cap stocks (sorted monthly), with the R^2_{oos} results presented in Table 2. The training, validation, and testing are conducted using the respective part of the data.

As shown in Table 2, on the entire out-of-sample set, increasing the number of layers in the neural network gradually improves the R^2_{oos}, peaking at NN3 and NN4 with a high of approximately 0.40%, followed by a decline. For the two specific out-of-sample sets (Top and Bottom), a similar trend is observed: performance peaks at NN3 and NN4, then decreases. In summary, the predictive performance of the single-layer neural network model (NN1) is relatively poor, and adding more layers does not necessarily lead to better results. Overall, our results are similar to those reported in Gu et al. (2020) and Cakici et al. (2023) which studied the risk premium predictability with machine learning, despite small differences in datasets and settings. The small R^2_{oos} values reported in all these studies indicate the weak predictability of the risk premium, aligning with the low signal-to-noise ratios in financial data as discussed in Shen and Xiu (2024). However, this evaluation only considers the forecasting of the risk premium, without accounting for conditional volatility.

To evaluate the forecasting of conditional volatility, we report the average log Gaussian likelihood in Table 3, using the realized $r_{i,t+1}$ alongside the predicted $\hat{\mu}_{i,t+1}$ and $\hat{\sigma}_{i,t+1}$ (parameters of the Gaussian distribution) across the entire out-of-sample set. A higher log likelihood value indicates better performance. As seen in Table 3, NN2, NN3, and NN4 deliver similar performance, while NN5 performs poorly. Interestingly, NN1 outperforms NN5 in terms of log likelihood, even though NN5 surpasses NN1 in R^2_{oos}, as shown in Table 2. In the remainder of this subsection, we qualitatively evaluate the forecasting of conditional volatility, focusing solely on the NN3 model for the sake of conciseness without loss of generality.

To comprehensively evaluate the forecasting of conditional volatility, we sort all stocks into deciles each month based on the predicted conditional volatilities $\hat{\sigma}_{i,t+1}$ in ascending order. We then calculate the standard deviation of realized excess returns $r_{i,t+1}$ within each decile for each month. Figure 1 illustrates the time-varying standard deviations of each decile across the entire twenty-year out-of-sample period, using predicted conditional volatilities from NN3. The line labeled as n represents the standard deviations in the n-th decile. As observed, in most months, the standard deviation of realized excess returns aligns with the decile rank of the predicted conditional volatilities. Overall, the levels of the ten lines are in ascending order, as expected. This suggests that the predictability of conditional volatility is significant and that our predictions are accurate.

Furthermore, we aggregate the twelve n-th deciles in each year and quantitatively analyze the standard deviation of realized excess returns annually. Using the predicted conditional volatilities from NN3, as displayed in Table 4, we compute the standard deviation of realized excess returns for each decile annually, where $n \in \{1, 2, \ldots, 10\}$. Table 4 illustrates that, in most years, there is a perfect match between the ranking of standard deviations of realized returns and the ranking of

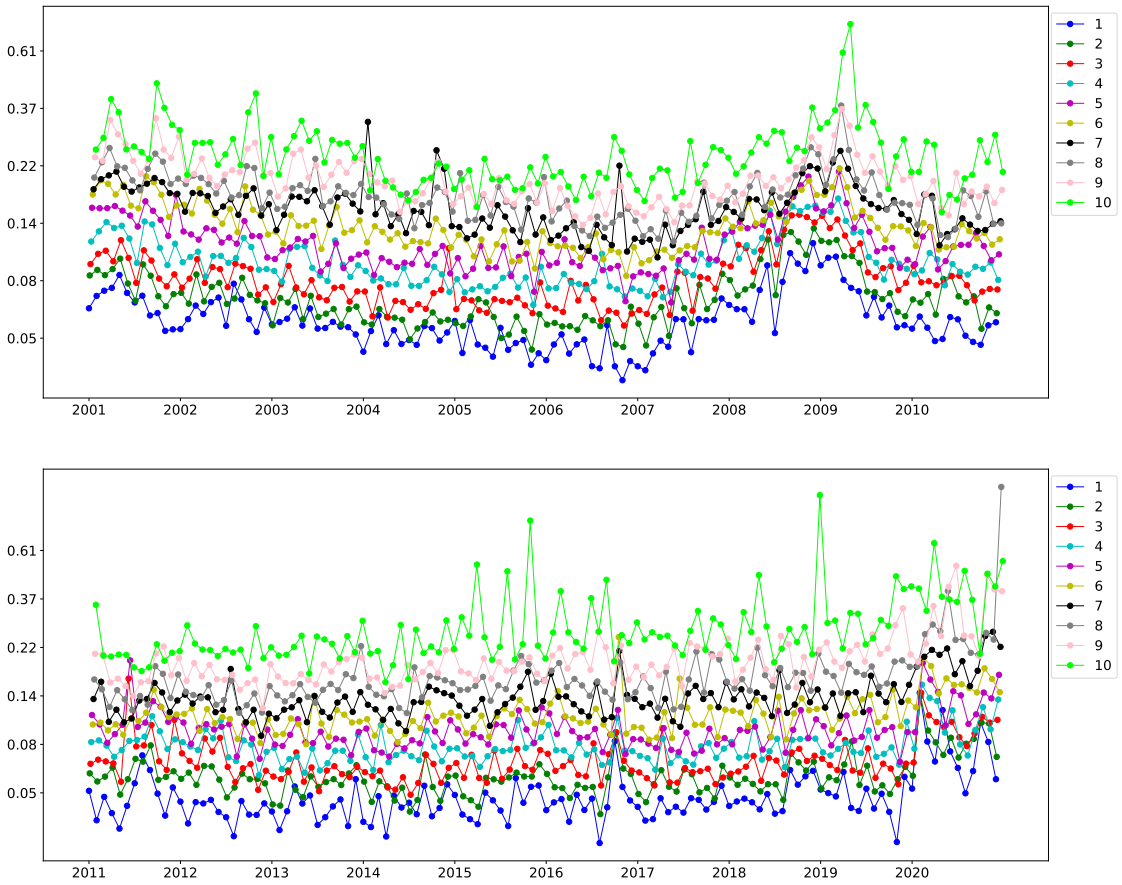


Figure 1: The standard deviation of realized excess returns $r_{i,t+1}$ within each decile for each month. The ten deciles are obtained by sorting all stocks each month based on the predicted conditional volatilities $\hat{\sigma}_{i,t+1}$ in ascending order. The line labeled as n represents the standard deviations in the n-th decile across the entire twenty-year out-of-sample period. As observed, in most months, the standard deviation of realized excess returns aligns with the decile rank of the predicted conditional volatilities. Overall, the levels of the ten lines are in ascending order, as expected.

Table 4: The annual standard deviation of realized excess returns for each decile.

	1	2	3	4	5	6	7	8	9	10
2001	0.072	0.088	0.103	0.129	0.160	0.188	0.209	0.245	0.296	0.353
2002	0.068	0.080	0.095	0.110	0.132	0.164	0.177	0.208	0.242	0.314
2003	0.059	0.067	0.080	0.101	0.111	0.137	0.165	0.188	0.227	0.283
2004	0.054	0.062	0.077	0.088	0.104	0.126	0.188	0.169*	0.195	0.212
2005	0.050	0.062	0.073	0.086	0.103	0.118	0.142	0.174	0.185	0.215
2006	0.045	0.057	0.068	0.084	0.098	0.112	0.135	0.146	0.165	0.215
2007	0.055	0.069	0.081	0.094	0.104	0.120	0.140	0.152	0.172	0.224
2008	0.090	0.113	0.135	0.151	0.170	0.177	0.196	0.218	0.233	0.295
2009	0.084	0.101	0.119	0.136	0.154	0.171	0.200	0.232	0.267	0.416
2010	0.064	0.082	0.095	0.107	0.124	0.134	0.153	0.166	0.190	0.245
2011	0.058	0.075	0.102	0.103	0.131	0.129*	0.145	0.164	0.198	0.235
2012	0.045	0.063	0.074	0.088	0.099	0.111	0.131	0.143	0.171	0.232
2013	0.051	0.061	0.068	0.079	0.093	0.108	0.125	0.156	0.180	0.233
2014	0.050	0.059	0.066	0.083	0.094	0.109	0.130	0.153	0.182	0.232
2015	0.050	0.064	0.075	0.088	0.104	0.116	0.137	0.171	0.190	0.383
2016	0.055	0.064	0.075	0.087	0.100	0.131	0.145	0.161	0.210	0.305
2017	0.045	0.058	0.066	0.076	0.086	0.110	0.130	0.165	0.205	0.263
2018	0.060	0.072	0.083	0.096	0.110	0.129	0.151	0.185	0.214	0.422
2019	0.052	0.068	0.077	0.090	0.105	0.127	0.144	0.180	0.239	0.330
2020	0.101	0.117	0.129	0.145	0.168	0.195	0.244	0.449	0.378*	0.463

Notes. We aggregate the twelve n-th deciles in each year, where $n \in \{1, 2, \ldots, 10\}$. The predicted conditional volatilities are from NN3. In most years, there is a perfect match between the ranking of standard deviations of realized returns and the ranking of predicted conditional volatilities, both in ascending order. The three exceptions are marked with $*$.

predicted conditional volatilities, both in ascending order. The three exceptions include the 8-th decile in 2004, the 6-th decile in 2011, and the 9-th decile in 2020.

4.3. Neural Network Double-Sorted Portfolios

In this subsection, we construct double-sorted portfolios based on risk premiums and conditional volatilities predicted by neural networks, which yield impressive Sharpe ratio results. Prior research, including Gu et al. (2020), Cakici et al. (2023), and Leippold et al. (2022), has used various machine learning models to forecast risk premiums or expected returns. These models have been applied to create single-sorted portfolios by ranking stocks into 10 deciles based on the predicted risk premiums. Subsequently, a zero-investment long-short portfolio was formed by taking long positions in the highest decile and short positions in the lowest decile, and the performance of this portfolio was analyzed.

The predicted risk premium and predicted conditional volatility are both of interest in our analysis. Following the methodology of Bali et al. (2016), we perform double sorting, or bivariate sorting, over the entire twenty-year out-of-sample period based on these two variables. Additionally, double sorting can be categorized into two types: independent and dependent. We also conduct single sorting based solely on predicted risk premiums for comparative analysis. In brief, when controlling for conditional volatility, the transition from single sorting to double sorting results in a significant improvement in the performance of the long-short portfolio.

4.3.1. Portfolio Performance

Each month, we employ both single/univariate sorting and double/bivariate sorting techniques—independent and dependent—to construct ten investment portfolios and one high-minus-low zero-investment portfolio. We utilize two weighting schemes: equal weights and market equity value weights. In double sorting, the predicted conditional volatility serves as the primary sorting variable or controlling variable. In both independent and dependent double sorting, stocks are divided into 10×10 groups, $G_{i,j}$, $1 \leq i \leq 10$, $1 \leq j \leq 10$, where i denotes the i-th group based on the predicted risk premiums and j denotes the j-th group based on the predicted conditional volatilities. The ten investment portfolios are constructed by taking long positions in the groups $G_{i,1:10}$. Specifically, within the i-th portfolio, we purchase stocks in $G_{i,j}$ (with equal weights or value weights) and assign equal weights across the ten groups $G_{i,1:10}$. The high-minus-low zero-investment portfolio is established by taking long positions in the 10th portfolio and short positions in the 1st portfolio.

Table 5 presents the performance of portfolios using equal weights across five models and three sorting methods. The realized average excess returns, denoted as Real, generally exhibit an ascending trend across the ten portfolios, with only three exceptions: the 8th portfolio in NN4 Independent Double, the 6th portfolio in NN5 Independent Double, and the 8th portfolio in NN5 Dependent Double. The realized average excess returns of all the single-sorted H-L portfolios (ranging from 2.32% to 2.75%) are typically slightly higher than those of the double-sorted H-L portfolios, whether independent (2.15% to 2.62%) or dependent (1.96% to 2.26%). However, the

Table 5: Performance of the neural network portfolios using equal weights.

	NN1 Single				NN1 Independent Double				NN1 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-2.81	-0.78	8.79	-0.31	-2.44	-0.51	6.53	-0.27	-2.27	-0.47	6.27	-0.26
2	-0.97	0.25	6.91	0.12	-0.96	0.23	6.02	0.13	-1.01	0.17	6.01	0.10
3	-0.27	0.54	5.87	0.32	-0.27	0.58	5.96	0.34	-0.40	0.33	5.95	0.19
4	0.19	0.72	5.39	0.46	0.19	0.74	5.95	0.43	0.07	0.73	5.99	0.43
5	0.57	0.82	5.06	0.56	0.57	0.76	5.73	0.46	0.48	0.86	5.86	0.51
6	0.91	0.99	5.08	0.67	0.91	0.98	5.80	0.58	0.87	1.01	5.83	0.60
7	1.27	1.13	5.09	0.77	1.27	1.16	5.80	0.69	1.27	1.18	5.79	0.71
8	1.68	1.30	5.22	0.86	1.68	1.33	5.72	0.80	1.72	1.28	5.69	0.78
9	2.26	1.47	5.68	0.90	2.26	1.48	5.82	0.88	2.29	1.51	5.84	0.90
High(H)	3.70	1.97	6.72	1.01	3.52	1.90	5.89	1.12	3.48	1.79	5.80	1.07
H-L	6.51	2.75	4.46	2.13	5.98	2.47	2.91	2.93	5.75	2.26	2.66	2.94
	NN2 Single				NN2 Independent Double				NN2 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-2.35	-0.81	9.06	-0.31	-2.00	-0.63	6.76	-0.32	-1.73	-0.45	6.19	-0.25
2	-0.67	0.28	7.19	0.13	-0.64	0.29	6.00	0.17	-0.68	0.19	6.02	0.11
3	-0.03	0.57	6.08	0.33	-0.03	0.54	5.97	0.32	-0.18	0.40	6.03	0.23
4	0.34	0.73	5.30	0.48	0.34	0.76	5.95	0.44	0.19	0.68	5.87	0.40
5	0.63	0.90	4.93	0.63	0.63	1.02	5.81	0.61	0.52	0.85	6.00	0.49
6	0.89	1.02	4.81	0.73	0.89	1.08	5.65	0.67	0.83	1.03	5.76	0.62
7	1.15	1.07	4.90	0.76	1.16	1.11	5.72	0.67	1.15	1.12	5.82	0.67
8	1.47	1.26	5.21	0.84	1.47	1.26	5.66	0.77	1.50	1.28	5.67	0.78
9	1.92	1.49	5.67	0.91	1.92	1.50	5.87	0.88	1.96	1.50	5.74	0.90
High(H)	3.18	1.90	6.90	0.95	2.98	1.90	6.03	1.09	2.95	1.78	5.96	1.04
H-L	5.53	2.71	4.77	1.97	5.02	2.62	3.43	2.65	4.69	2.23	2.62	2.95
	NN3 Single				NN3 Independent Double				NN3 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-1.93	-0.67	9.04	-0.26	-1.62	-0.45	6.49	-0.24	-1.36	-0.40	6.12	-0.23
2	-0.44	0.31	7.35	0.15	-0.42	0.28	6.19	0.16	-0.42	0.29	6.21	0.16
3	0.13	0.62	6.13	0.35	0.13	0.62	5.93	0.36	0.00	0.47	6.05	0.27
4	0.46	0.75	5.38	0.48	0.46	0.78	5.95	0.45	0.32	0.64	5.94	0.37
5	0.70	0.85	4.98	0.59	0.70	0.88	5.92	0.52	0.59	0.88	5.87	0.52
6	0.91	0.97	4.80	0.70	0.91	0.97	5.65	0.59	0.84	1.06	5.81	0.63
7	1.12	1.10	4.89	0.78	1.12	1.16	5.87	0.68	1.09	1.10	5.70	0.67
8	1.36	1.27	5.17	0.85	1.36	1.33	5.90	0.78	1.37	1.16	5.75	0.70
9	1.70	1.42	5.63	0.87	1.70	1.35	5.78	0.81	1.73	1.45	5.82	0.86
High(H)	2.63	1.78	6.77	0.91	2.49	1.84	5.93	1.08	2.45	1.75	5.85	1.04
H-L	4.56	2.45	4.88	1.74	4.13	2.40	3.37	2.47	3.80	2.15	2.70	2.76
	NN4 Single				NN4 Independent Double				NN4 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-1.50	-0.61	8.98	-0.24	-1.28	-0.32	6.61	-0.17	-1.03	-0.31	5.99	-0.18
2	-0.34	0.31	7.24	0.15	-0.32	0.30	6.07	0.17	-0.31	0.19	6.05	0.11
3	0.14	0.60	6.13	0.34	0.14	0.59	5.82	0.35	0.04	0.52	5.96	0.30
4	0.43	0.74	5.37	0.48	0.43	0.78	5.97	0.45	0.31	0.69	5.95	0.40
5	0.64	0.88	5.01	0.61	0.64	0.99	6.06	0.57	0.55	0.82	5.88	0.48
6	0.83	0.98	4.96	0.68	0.83	1.07	5.93	0.63	0.77	1.06	5.95	0.61
7	1.02	1.16	4.98	0.81	1.02	1.25	5.79	0.75	0.99	1.19	5.91	0.70
8	1.23	1.20	5.24	0.80	1.23	1.13	5.81	0.67	1.24	1.19	5.82	0.71
9	1.55	1.45	5.66	0.89	1.55	1.56	6.26	0.86	1.56	1.39	5.80	0.83
High(H)	2.38	1.71	6.59	0.90	2.26	1.71	5.99	0.99	2.21	1.66	5.79	0.99
H-L	3.88	2.32	4.96	1.62	3.57	2.17	3.78	1.99	3.25	1.96	2.61	2.61
	NN5 Single				NN5 Independent Double				NN5 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-1.06	-0.64	8.90	-0.25	-0.90	-0.26	6.79	-0.13	-0.72	-0.34	6.10	-0.19
2	-0.15	0.33	7.25	0.16	-0.14	0.33	6.10	0.19	-0.13	0.23	6.16	0.13
3	0.21	0.61	6.24	0.34	0.21	0.59	5.87	0.35	0.15	0.55	6.01	0.31
4	0.44	0.81	5.53	0.51	0.44	0.83	5.94	0.48	0.36	0.68	5.80	0.40
5	0.61	0.99	5.24	0.66	0.61	1.04	6.00	0.60	0.55	0.87	5.87	0.51
6	0.76	0.99	5.03	0.68	0.76	0.98	5.82	0.58	0.72	0.98	5.82	0.58
7	0.91	1.10	5.04	0.76	0.91	1.13	5.86	0.67	0.89	1.19	5.86	0.70
8	1.09	1.17	5.23	0.77	1.09	1.18	5.89	0.70	1.08	1.16	5.79	0.70
9	1.32	1.36	5.33	0.88	1.32	1.37	5.69	0.83	1.31	1.40	5.86	0.83
High(H)	1.85	1.69	6.35	0.92	1.79	1.76	5.93	1.03	1.76	1.68	5.79	1.01
H-L	2.91	2.33	5.04	1.60	2.70	2.15	3.48	2.14	2.48	2.03	2.60	2.71

Notes. In double sorting, the predicted conditional volatility is the primary sorting variable or controlling variable. Pred represents the predicted average excess return/risk premium of the portfolio (monthly), Real represents the realized average excess return (monthly), Std is the realized excess returns' standard deviation (monthly), and SR is the Sharpe ratio (annually).

Table 6: Performance of the neural network portfolios using value weights.

	NN1 Single				NN1 Independent Double				NN1 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-2.44	-0.36	9.14	-0.14	-2.37	-0.17	8.30	-0.07	-2.21	-0.22	7.82	-0.10
2	-0.92	0.19	6.58	0.10	-0.94	0.21	7.25	0.10	-1.01	0.22	7.34	0.10
3	-0.25	0.31	5.17	0.20	-0.27	0.44	6.87	0.22	-0.40	0.21	6.83	0.11
4	0.20	0.45	4.48	0.35	0.19	0.61	6.74	0.32	0.07	0.50	6.97	0.25
5	0.57	0.64	4.39	0.50	0.57	0.64	6.71	0.33	0.47	0.75	6.78	0.38
6	0.91	0.71	4.37	0.56	0.91	0.70	6.78	0.36	0.86	0.81	6.67	0.42
7	1.27	0.77	4.31	0.62	1.27	0.95	6.47	0.51	1.26	0.92	6.58	0.48
8	1.68	1.03	4.75	0.75	1.68	1.14	6.73	0.59	1.71	1.16	6.52	0.61
9	2.23	1.01	5.16	0.68	2.25	1.13	6.54	0.60	2.29	1.18	6.65	0.62
High(H)	3.36	1.16	6.47	0.62	3.44	1.32	6.63	0.69	3.42	1.11	6.58	0.58
H-L	5.80	1.52	5.59	0.94	5.83	1.53	4.02	1.32	5.63	1.33	3.67	1.25
	NN2 Single				NN2 Independent Double				NN2 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-2.05	-0.37	9.91	-0.13	-1.93	-0.45	8.23	-0.19	-1.67	-0.15	7.89	-0.07
2	-0.60	0.39	7.05	0.19	-0.62	0.36	7.15	0.18	-0.68	0.29	7.16	0.14
3	-0.02	0.35	5.34	0.23	-0.03	0.42	7.04	0.20	-0.18	0.32	6.92	0.16
4	0.35	0.52	4.44	0.41	0.34	0.61	6.62	0.32	0.19	0.59	6.77	0.30
5	0.63	0.65	4.42	0.51	0.63	0.79	6.85	0.40	0.52	0.66	6.87	0.33
6	0.89	0.76	4.45	0.59	0.89	0.99	6.71	0.51	0.83	1.02	6.84	0.52
7	1.15	0.79	4.24	0.65	1.15	0.92	6.55	0.49	1.14	0.89	6.48	0.48
8	1.46	0.78	4.73	0.57	1.47	1.05	6.50	0.56	1.50	1.04	6.49	0.56
9	1.89	0.98	5.24	0.65	1.91	1.14	6.49	0.61	1.96	1.16	6.44	0.62
High(H)	2.88	1.08	7.02	0.53	2.94	1.31	7.09	0.64	2.91	1.06	6.80	0.54
H-L	4.92	1.45	6.28	0.80	4.91	1.83	4.42	1.44	4.57	1.22	3.72	1.13
	NN3 Single				NN3 Independent Double				NN3 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-1.66	-0.17	9.83	-0.06	-1.56	-0.19	8.13	-0.08	-1.29	-0.11	7.74	-0.05
2	-0.39	0.17	7.31	0.08	-0.40	0.19	7.35	0.09	-0.42	0.29	7.55	0.13
3	0.14	0.38	5.50	0.24	0.13	0.43	6.89	0.22	0.00	0.36	7.06	0.18
4	0.47	0.44	4.72	0.32	0.46	0.60	6.88	0.30	0.32	0.52	6.98	0.26
5	0.70	0.77	4.32	0.62	0.70	0.85	6.96	0.42	0.59	0.68	6.68	0.35
6	0.91	0.66	4.28	0.53	0.91	0.85	6.60	0.45	0.84	0.98	6.75	0.50
7	1.11	0.79	4.28	0.64	1.12	0.99	6.75	0.51	1.09	0.86	6.50	0.46
8	1.35	0.88	4.69	0.65	1.36	1.13	6.71	0.58	1.37	1.03	6.65	0.53
9	1.68	0.86	5.37	0.56	1.69	0.95	6.64	0.50	1.72	1.19	6.59	0.62
High(H)	2.42	1.21	6.84	0.61	2.44	1.35	6.75	0.69	2.40	1.09	6.76	0.56
H-L	4.08	1.39	6.22	0.77	4.03	1.65	4.30	1.33	3.70	1.20	4.04	1.03
	NN4 Single				NN4 Independent Double				NN4 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-1.31	-0.26	10.06	-0.09	-1.23	-0.23	8.28	-0.10	-0.98	-0.11	7.70	-0.05
2	-0.30	0.31	7.81	0.14	-0.31	0.32	7.38	0.15	-0.31	0.16	7.14	0.08
3	0.16	0.35	5.71	0.21	0.15	0.43	6.96	0.21	0.04	0.54	7.15	0.26
4	0.43	0.41	4.57	0.31	0.43	0.60	6.78	0.31	0.31	0.52	6.91	0.26
5	0.64	0.57	4.29	0.46	0.64	0.71	6.83	0.36	0.55	0.64	6.87	0.32
6	0.83	0.75	4.42	0.59	0.83	0.92	6.55	0.49	0.77	0.91	6.82	0.46
7	1.01	0.65	4.28	0.53	1.01	1.02	6.67	0.53	0.99	0.85	6.65	0.44
8	1.23	1.04	4.54	0.79	1.23	0.92	6.50	0.49	1.24	1.12	6.56	0.59
9	1.53	1.05	5.19	0.70	1.54	1.23	6.68	0.64	1.56	0.99	6.50	0.53
High(H)	2.19	1.00	6.25	0.56	2.21	1.32	7.17	0.64	2.16	1.09	6.57	0.58
H-L	3.49	1.26	6.60	0.66	3.47	1.70	5.74	1.03	3.14	1.20	3.82	1.09
	NN5 Single				NN5 Independent Double				NN5 Dependent Double			
	Pred	Real	Std	SR	Pred	Real	Std	SR	Pred	Real	Std	SR
Low(L)	-0.89	-0.22	9.55	-0.08	-0.85	-0.05	8.35	-0.02	-0.68	-0.16	7.62	-0.07
2	-0.12	0.37	7.59	0.17	-0.13	0.25	7.32	0.12	-0.13	0.22	7.36	0.10
3	0.22	0.18	5.73	0.11	0.21	0.36	6.84	0.18	0.15	0.38	7.02	0.19
4	0.44	0.52	4.93	0.37	0.44	0.51	6.82	0.26	0.36	0.58	6.95	0.29
5	0.61	0.62	4.56	0.47	0.61	0.79	6.95	0.39	0.55	0.48	6.94	0.24
6	0.76	0.65	4.37	0.51	0.76	0.93	6.72	0.48	0.72	0.86	6.78	0.44
7	0.91	0.80	4.16	0.67	0.91	1.00	6.67	0.52	0.89	0.93	6.55	0.49
8	1.08	0.74	4.39	0.58	1.09	0.94	6.59	0.49	1.08	0.90	6.48	0.48
9	1.31	0.93	5.00	0.64	1.31	1.00	6.72	0.52	1.31	1.15	6.66	0.60
High(H)	1.74	1.12	5.92	0.65	1.76	1.43	6.93	0.71	1.73	1.14	6.66	0.59
H-L	2.63	1.34	6.63	0.70	2.63	1.62	4.96	1.13	2.41	1.29	3.87	1.16

Notes. In double sorting, the predicted conditional volatility is the primary sorting variable or controlling variable. Pred represents the predicted average excess return/risk premium of the portfolio (monthly), Real represents the realized average excess return (monthly), Std is the realized excess returns' standard deviation (monthly), and SR is the Sharpe ratio (annually).

standard deviations, denoted as Std, of single-sorted H-L portfolios (4.46% to 5.04%) are significantly larger than those of double-sorted H-L portfolios, whether independent (2.91% to 3.78%) or dependent (2.60% to 2.70%). The standard deviations of dependent double-sorted H-L portfolios remain consistent across the five models, ranging from 2.60% to 2.70%, which annualizes to 9.01% to 9.35%. Consequently, the annualized Sharpe ratios (SR) for the three sorting methods range as follows: 1.60 to 2.13 for single sorting, 1.99 to 2.93 for independent double sorting, and 2.61 to 2.95 for dependent double sorting. The Sharpe ratios of dependent double-sorted H-L portfolios are approximately 1.0 higher compared to those of single-sorted.

In Table 5, across the five models, dependent double sorting consistently yields higher annualized Sharpe ratios compared to independent double sorting. Specifically, Sharpe ratios from independent sorting range from 1.99 to 2.93, whereas dependent sorting yields Sharpe ratios between 2.61 and 2.95, demonstrating markedly superior performance. Notably, the highest Sharpe ratio of 2.95 is achieved by NN2 with dependent double sorting, with an average monthly excess return of 2.23% (annualized 26.76%) and a monthly volatility of 2.62% (annualized 9.08%). All results are out-of-sample.

Table 6 details the performance of portfolios using value weights across five models and three sorting methods. Similar to the case with equal weights, the realized average excess returns, labeled as Real, exhibit a generally ascending trend across the ten portfolios. Comparing single sorting with double sorting reveals similar patterns to those observed in equal-weighted portfolios in Table 5. For instance, the standard deviations of single-sorted H-L portfolios, ranging from 5.59% to 6.63%, are notably higher than those of independent double-sorted H-L portfolios (4.02% to 5.74%) and dependent ones (3.67% to 4.04%). Similarly, the Sharpe ratios of single-sorted H-L portfolios (0.66 to 0.94) are significantly lower than those of independent double-sorted H-L portfolios (1.03 to 1.44) and dependent ones (1.03 to 1.25). The Sharpe ratios of independent double-sorted H-L portfolios are approximately 0.4 higher compared to single-sorted H-L portfolios. Notably, independent double-sorted H-L portfolios also achieve significantly higher realized average excess returns (1.53% to 1.83%) compared to single-sorted H-L portfolios (1.26% to 1.52%), a contrast to the findings for equal weights in Table 5.

In Table 6, only for NN2 and NN3, independent double sorting yields significantly higher annualized Sharpe ratios compared to dependent double sorting (1.44 vs. 1.13 for NN2 and 1.33 vs. 1.03 for NN3). For other models, there are no significant differences. Notably, NN2 achieves the highest Sharpe ratio of 1.44 with independent double sorting, which also features the highest average monthly excess return of 1.83% (annualized 21.96%) and a monthly volatility of 4.42% (annualized 15.31%). All results are out-of-sample.

Comparatively, in Cakici et al. (2023), the highest Sharpe ratio achieved by various machine learning models on the U.S. market using the single sorting strategy is 1.97 (with equal weights) and 1.09 (with value weights), which are close to our ratios of 2.13 and 0.94, respectively. The data and settings in Cakici et al. (2023) are similar to ours. However, our results are not directly comparable to those in Gu et al. (2020) due to differences in datasets and settings. Nonetheless,

it is noteworthy that Gu et al. (2020) reported a highest Sharpe ratio of 2.45 (with equal weights) and 1.35 (with value weights) using machine learning models and the single sorting strategy on the U.S. market. These results all suggest that double sorting consistently yields significantly higher Sharpe ratios than single sorting, with an approximate increase of 1.0 when using equal weights and 0.4 when using value weights. Therefore, we provide investors with an innovative approach to risk mitigation and performance improvement in zero-investment scenarios.

4.3.2. Alternative Metrics

Simultaneously, we examine alternative metrics to evaluate the performance of the aforementioned portfolios, with a particular focus on the zero-investment long-short portfolios. Table 7 displays essential metrics, including maximum drawdown, portfolio turnover, and risk-adjusted performance with respect to the well-known pricing factors. This analysis encompasses both equal-weighted and value-weighted portfolios derived from the five neural network models and three sorting methods. The top two panels provide the maximum drawdown (Max DD), the most extreme negative monthly return (Max 1M loss), and the average monthly turnover. The bottom two panels present the realized mean return (in percentage), as well as the alpha, R^2, and information ratio (IR) with respect to the Fama-French five factors plus the momentum factor.

We first explore alternative risk indicators beyond standard deviation, firstly focusing on maximum drawdown, which indicates the potential maximum loss of a portfolio. The maximum drawdown of a portfolio is defined as

$$\text{Max DD} = \max_{0 \le t_1 \le t_2 \le T} \frac{Y_{t_1} - Y_{t_2}}{1 + Y_{t_1}}, \tag{10}$$

where Y_t is the cumulative return from time 0 to t. Across the five neural network structures, Table 7 shows that portfolios constructed using the double sorting method consistently exhibit significantly lower maximum drawdowns compared to those employing the single sorting method. In the equal-weighted portfolios, the NN2 model with the dependent double sorting method achieves the most favorable outcome, with the lowest maximum drawdown of 6.02%. Similarly, in the value-weighted portfolios, the NN2 model with the independent double sorting method displays the lowest maximum drawdown at 14.11%. The single sorting method yields maximum drawdowns exceeding 40% across all combinations of neural networks and weighting strategies.

We also assess the most extreme monthly loss as an additional risk indicator. In Table 7, the maximum monthly losses for double-sorted long-short portfolios consistently remain lower than those for single-sorted portfolios, further highlighting their robust performance. In the equal-weighted portfolios, the NN3 model with the dependent double sorting method achieves the best result, with the lowest maximum monthly loss of 3.33%. For the value-weighted portfolios, the NN2 model with the independent double sorting method leads with the smallest extreme monthly loss of 9.34%. The single sorting method, in contrast, results in maximum monthly losses exceeding 20% across all neural network and weighting strategy combinations. This analysis underscores the superior risk management capabilities of the double-sorting strategy incorporating predicted

Table 7: Drawdowns, turnover, and risk-adjusted performance of neural network long-short portfolios.

	NN1 sin	NN1 ind	NN1 dep	NN2 sin	NN2 ind	NN2 dep	NN3 sin	NN3 ind	NN3 dep	NN4 sin	NN4 ind	NN4 dep	NN5 sin	NN5 ind	NN5 dep
Drawdowns and turnover (equal-weighted)															
Max DD (%)	41.30	8.12	7.26	49.48	14.59	6.02	49.73	16.96	8.41	55.75	18.23	13.39	51.53	22.93	9.20
Max 1M loss (%)	21.70	8.12	4.54	24.10	11.54	4.05	24.41	6.93	3.33	23.29	9.34	4.92	23.05	6.65	3.95
Turnover (%)	226.80	272.80	251.70	210.82	264.33	242.58	198.42	255.11	229.82	197.53	252.18	227.02	208.27	257.76	234.01
Drawdowns and turnover (value-weighted)															
Max DD (%)	42.71	20.98	19.81	50.57	14.11	23.89	49.22	15.45	26.78	60.00	16.63	29.30	60.46	26.12	25.67
Max 1M loss (%)	21.82	9.91	13.90	27.22	9.34	11.36	26.08	9.56	13.45	33.64	10.76	12.97	30.82	11.74	9.52
Turnover (%)	268.04	285.48	277.10	247.94	278.13	272.35	236.96	271.93	263.79	239.29	270.98	265.04	245.22	274.22	266.40
Risk-adjusted performance (equal-weighted)															
Mean return	2.75	2.47	2.26	2.71	2.62	2.23	2.45	2.40	2.15	2.32	2.17	1.96	2.33	2.15	2.03
FF5+Mom α	2.73	2.30	2.19	2.72	2.38	2.12	2.43	2.08	2.01	2.29	1.77	1.80	2.32	1.88	1.85
$t(\alpha)$	10.42	12.30	12.73	9.88	10.87	12.74	8.68	9.93	11.81	8.22	7.50	11.00	8.18	8.52	11.33
R^2	28.41	14.81	13.21	30.78	15.65	16.40	31.70	19.83	17.47	34.44	19.33	18.70	34.29	16.29	17.49
IR	0.72	0.85	0.88	0.69	0.75	0.88	0.60	0.69	0.82	0.57	0.52	0.76	0.57	0.59	0.79
Risk-adjusted performance (value-weighted)															
Mean return	1.52	1.53	1.33	1.45	1.83	1.22	1.39	1.65	1.20	1.26	1.70	1.20	1.34	1.62	1.29
FF5+Mom α	1.47	1.44	1.33	1.32	1.50	1.08	1.29	1.31	1.06	1.19	1.09	1.06	1.32	1.29	1.08
$t(\alpha)$	4.61	5.66	5.63	3.80	5.37	4.67	3.74	4.97	4.22	3.41	3.02	4.60	3.72	4.08	4.46
R^2	32.49	16.41	13.71	36.43	17.05	19.32	36.50	21.48	20.26	42.47	17.26	23.69	40.08	15.30	19.19
IR	0.32	0.39	0.39	0.26	0.37	0.32	0.26	0.34	0.29	0.24	0.21	0.32	0.26	0.28	0.31

Notes. The top two panels provide the maximum drawdown (Max DD), the most extreme negative monthly return (Max 1M loss), and the average monthly turnover for both equal-weighted and value-weighted portfolios. The bottom two panels present the realized mean return (in percentage), as well as the alpha, R^2, and information ratio (IR) with respect to the Fama-French five factors plus the momentum factor, for both equal-weighted and value-weighted portfolios. "sin", "ind", and "dep" represent single sorting, independent double sorting, and dependent double sorting, respectively.

conditional volatilities, enhancing the overall resilience of the long-short portfolios.

Further, we examine the turnover rates of the portfolios. Turnover rate represents the frequency at which assets in the portfolio are bought and sold. We adopt the definition of turnover rate in Gu et al. (2020):

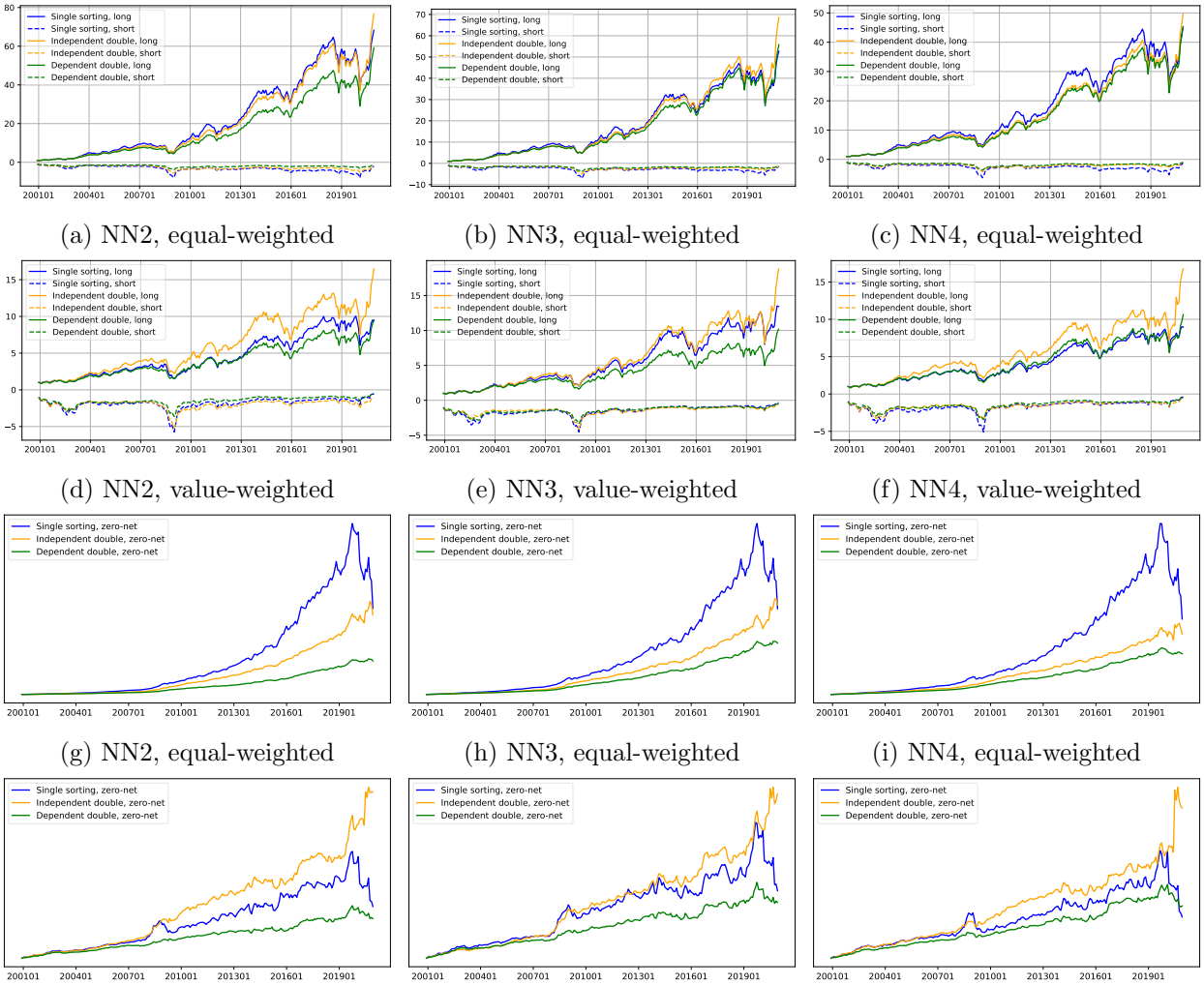
$$\text{Turnover} = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i}\left|w_{i,t+1} - \frac{w_{i,t}\left(1+r_{i,t+1}\right)}{1+\sum_{j} w_{j,t} r_{j,t+1}}\right|\right), \tag{11}$$

where $w_{i,t}$ is the weight of stock i in the portfolio at time t and $r_{i,t+1}$ is its return from t to $t+1$. From Table 7, in both equal-weighted portfolios and value-weighted portfolios, the single sorting strategy always exhibits lower turnover rates than the two double sorting methods, meeting our expectation. Comparing the two double sorting methods, dependent double sorting consistently results in lower turnover rates than independent double sorting.

To comprehensively assess the performance of our neural network portfolios, we employ a rigorous analysis of risk-adjusted performance utilizing factor pricing models. We utilize the Fama-French five-factor model (Fama and French, 2015), augmented with the momentum factor, as our benchmark. This model can capture the effects of market risk, size, value, profitability, investment, and momentum on portfolio performance. Table 7 presents key performance metrics for both single-sorted and double-sorted long-short portfolios. These metrics include mean returns, alphas, t-values of alpha, R^2, and information ratios (IR) with respect to the six factors. Notably, nearly all equal-weighted portfolios achieve mean returns exceeding 2.0%, while all value-weighted portfolios generate mean returns greater than 1.2% (both monthly). These mean returns are consistent with those reported in Tables 5 and 6.

Alpha represents the excess returns relative to the risk factors. A significant positive alpha indicates that the portfolio generates excess returns that cannot be explained by the exposure to common risk factors. Table 7 shows that our portfolios exhibit substantially positive alpha values, with most equal-weighted portfolios achieving alphas greater than 1.8% and all value-weighted portfolios exceeding 1.06%. These alphas are supported by statistically significant t-values, with most equal-weighted portfolios having t-values greater than 8 and all value-weighted portfolios exceeding 3. When comparing mean returns with alphas across the five models, three sorting methods, and two weighting strategies, we find that alpha is only slightly lower than the mean return.

The R^2 values reflect the explanatory power of the six-factor model, showing that double-sorted portfolios have significantly lower R^2 values (13.21% to 19.83%) compared to single-sorted portfolios (28.41% to 34.44%) for equal-weighted portfolios. A similar trend is observed for value-weighted portfolios. Across models and sorting methods, the information ratio (IR) typically ranges from 0.52 to 0.88 for equal-weighted portfolios and from 0.21 to 0.39 for value-weighted portfolios. Notably, double-sorted portfolios generally exhibit a moderately higher IR compared to single-sorted portfolios.



(a) NN2, equal-weighted (b) NN3, equal-weighted (c) NN4, equal-weighted

(d) NN2, value-weighted (e) NN3, value-weighted (f) NN4, value-weighted

(g) NN2, equal-weighted (h) NN3, equal-weighted (i) NN4, equal-weighted

(j) NN2, value-weighted (k) NN3, value-weighted (l) NN4, value-weighted

Figure 2: The cumulative return curves of both equal-weighted and value-weighted zero-investment long-short portfolios constructed using the NN2, NN3, and NN4 models. For brevity, the results of NN1 and NN5 have been omitted. The lines of different colors denote portfolios generated with different sorting strategies. In (a)–(f), solid lines represent long positions, while dashed lines represent short positions (the symmetric version along the horizontal axis). In (g)–(l), solid lines represent zero-investment portfolios themselves.

4.3.3. Cumulative Returns

In this section, we analyze the cumulative returns of the aforementioned zero-investment long-short portfolios constructed using the neural networks. Figure 2 presents the results of both equal-weighted and value-weighted portfolios from the NN2, NN3, and NN4 models, with NN1 and NN5 excluded for brevity. The long positions (solid lines) and short positions (dashed lines) of the zero-investment portfolios are separately depicted in (a)–(f), while (g)–(l) display the zero-investment portfolios themselves.

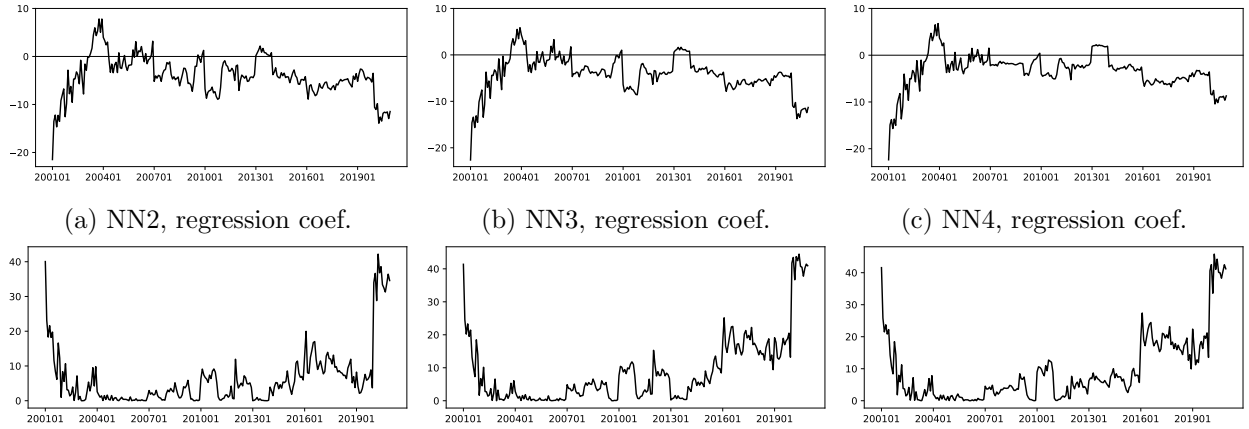
When examining the long positions and short positions separately in the subfigures (a)–(f), there are no discernible visual differences in the solid lines (representing long positions) across different sorting methods. The three solid lines exhibit similar fluctuation patterns in each subfigure. For the dashed lines representing short positions, our results align with those of Gu et al. (2020), showing that the cumulative return curves of short positions remain relatively flat from 2001 to 2020. However, a closer analysis reveals that the short positions associated with single sorting display greater volatility and more pronounced drawdowns, with the most significant drawdowns corresponding to the 2008 financial crisis and the 2020 COVID-19 crisis.

To further investigate and compare different sorting methods, we analyze the cumulative returns of zero-investment long-short portfolios, as illustrated in subfigures (g)–(l) of Figure 2. In the equal-weighted portfolios, single sorting generally yields higher cumulative returns in most of the time, although it experiences a severe loss during the COVID-19 crisis in 2020. Similarly, value-weighted portfolios also suffer from significant losses with single sorting, but independent double sorting results in the highest return curves. Both independent and dependent double sorting demonstrate more robust performance against extreme market conditions. The same observation is evident during the 2008 financial crisis. Throughout most periods, double-sorted zero-investment portfolios achieve stable growth with lower volatility. This stability is consistent with the significantly higher Sharpe ratios of the double-sorted zero-investment portfolios, as reported in Tables 5 and 6.

To elucidate why Sharpe ratios increase from single sorting to double sorting, consider the following analysis. The primary purpose of double sorting is to control for one variable while independently constructing the long-short portfolio based on another variable. If risk premium and conditional volatility were independent in cross-sections, we would expect no significant difference in portfolio performance between single and double sorting. However, the above results suggest that this is not the case, revealing a dependence between risk premium and conditional volatility. We will explore this relationship further in the subsequent analysis.

4.4. Relationship between Risk Premium and Conditional Volatility

As discussed in Section 2.3, existing studies have produced conflicting results regarding the relationship between risk premium and conditional volatility. In this section, we investigate their relationship using deep learning estimates based on predictions from the testing/out-of-sample set. Specifically, we perform ordinary linear regression, treating risk premium as the dependent variable



(a) NN2, regression coef. (b) NN3, regression coef. (c) NN4, regression coef.

(d) NN2, R squared (e) NN3, R squared (f) NN4, R squared

Figure 3: The regression coefficients (in percentage) and R^2 values plotted against the corresponding months. In each month, we regress the predicted risk premiums on the predicted conditional volatilities. For brevity, the results for NN1 and NN5 are not shown. The coefficients are predominantly negative throughout most of the time range.

and conditional volatility as the independent variable, on a monthly basis:

$$\hat{\mu}_{i,t+1} = \gamma_{0,t+1} + \gamma_{1,t+1}\hat{\sigma}_{i,t+1} + \varepsilon_{i,t+1}, \quad \text{for each } t. \tag{12}$$

We calculate the estimated regression coefficients and R^2 values for every month, which are plotted against the corresponding months in Figure 3. The results indicate that all five models yield similar outcomes, though for brevity, the results for NN1 and NN5 are not shown. Throughout most of the period, the regression coefficients are negative, with positive coefficients primarily occurring in 2003 and 2013. The R^2 values exceed 5% during most of the period and surpass 10% during 2001 and from 2016 to 2020. These suggest a negative relationship between risk premium and conditional volatility in cross-sections.

We also perform a statistical descriptive analysis of the regression coefficients, their corresponding t-values, and the R^2 values. Summary statistics of them across all out-of-sample months are presented in Table 8. The table reveals that, on average (as shown in the mean and median columns), the regression coefficients from NN1–NN5 suggest a negative relationship between risk premium and conditional volatility. The associated t-values (< -10) further affirm the statistical significance of this relationship. Additionally, the 0.75-quantile column indicates that, in most months, the coefficients are negative and statistically significant (t-values < -6).

When constructing long-short portfolios, the implication of the negative relationship between risk premium and conditional volatility is intuitive. In single sorting based on risk premium, stocks in short positions typically exhibit low risk premiums and, consequently, high conditional volatilities, whereas stocks in long positions demonstrate the opposite characteristics. However, high volatility is generally undesirable in portfolio construction. In double sorting, the volatilities of stocks in both long and short positions are balanced, approximating the average volatility of all

Table 8: Summary statistics of the regression coefficients, t-values, and R^2 values across all out-of-sample months.

	mean	std	min	.25 quantile	median	.75 quantile	max
Regression coefficient (in percentage)							
NN1	-3.97	4.14	-21.51	-6.36	-3.85	-1.52	6.55
NN2	-4.01	4.12	-21.48	-5.95	-4.22	-1.71	7.82
NN3	-4.09	3.96	-22.63	-5.71	-3.88	-1.90	5.86
NN4	-3.31	3.73	-22.37	-5.00	-2.93	-1.63	6.77
NN5	-3.71	3.53	-21.76	-5.78	-3.49	-1.97	6.98
t-value (negative only)							
NN1	-13.10	10.03	-62.56	-17.38	-10.91	-6.01	-0.40
NN2	-16.56	11.52	-65.11	-20.10	-14.96	-8.84	-0.24
NN3	-19.96	13.14	-66.93	-27.20	-18.31	-10.71	-0.51
NN4	-20.04	13.33	-67.21	-28.87	-17.31	-10.38	-0.59
NN5	-23.67	13.55	-66.60	-33.00	-22.80	-12.88	-0.69
R^2 (in percentage)							
NN1	4.56	6.41	0.00	0.62	1.99	5.81	38.14
NN2	6.39	8.37	0.00	0.86	3.93	8.24	42.20
NN3	8.97	10.18	0.01	1.37	5.49	13.88	44.47
NN4	9.34	10.23	0.00	1.75	5.86	14.41	45.81
NN5	11.51	11.33	0.00	1.94	8.72	17.16	42.65

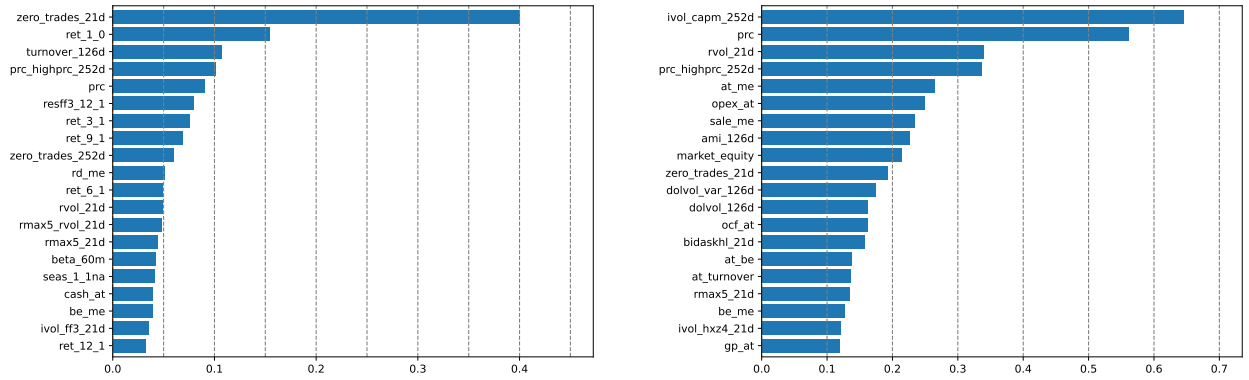
Notes. In each month, we regress the predicted risk premiums on the predicted conditional volatilities. In the middle panel, only negative t-values are collected and used for calculation. Most t-values are smaller than -6, indicating that the negative regression coefficients are significant. This table corresponds to Figure 3.

stocks within a given month. This balancing effect likely contributes to the superior Sharpe ratios observed in double-sorted long-short portfolios, compared to single-sorted ones.

4.5. Which Characteristics Matter?

Our objective is to identify the variables that significantly impact forecasting accuracy. To achieve this, we first calculate the variable importance (VI) for each variable and rank all characteristics based on their VI. Given that our model comprises two components—predicting risk premiums and predicting conditional volatilities—we evaluate the VI of variables for each component separately. The VI is computed using the in-sample set and averaged over the training sets corresponding to all rolling windows. Subsequently, we analyze these variables by examining their marginal effects as well as their interaction effects with other variables.

For risk premium prediction, we use the measurement approach proposed by Kelly et al. (2019) to assess VI. To determine the importance of the j-th variable, we set all values of this variable to zero while keeping the values of other variables unchanged. The importance measure is then quantified by the reduction in R^2 resulting from this modification. A larger reduction in R^2 indicates that the variable plays a more critical role in predicting risk premiums. For conditional volatility



(a) For risk premium prediction

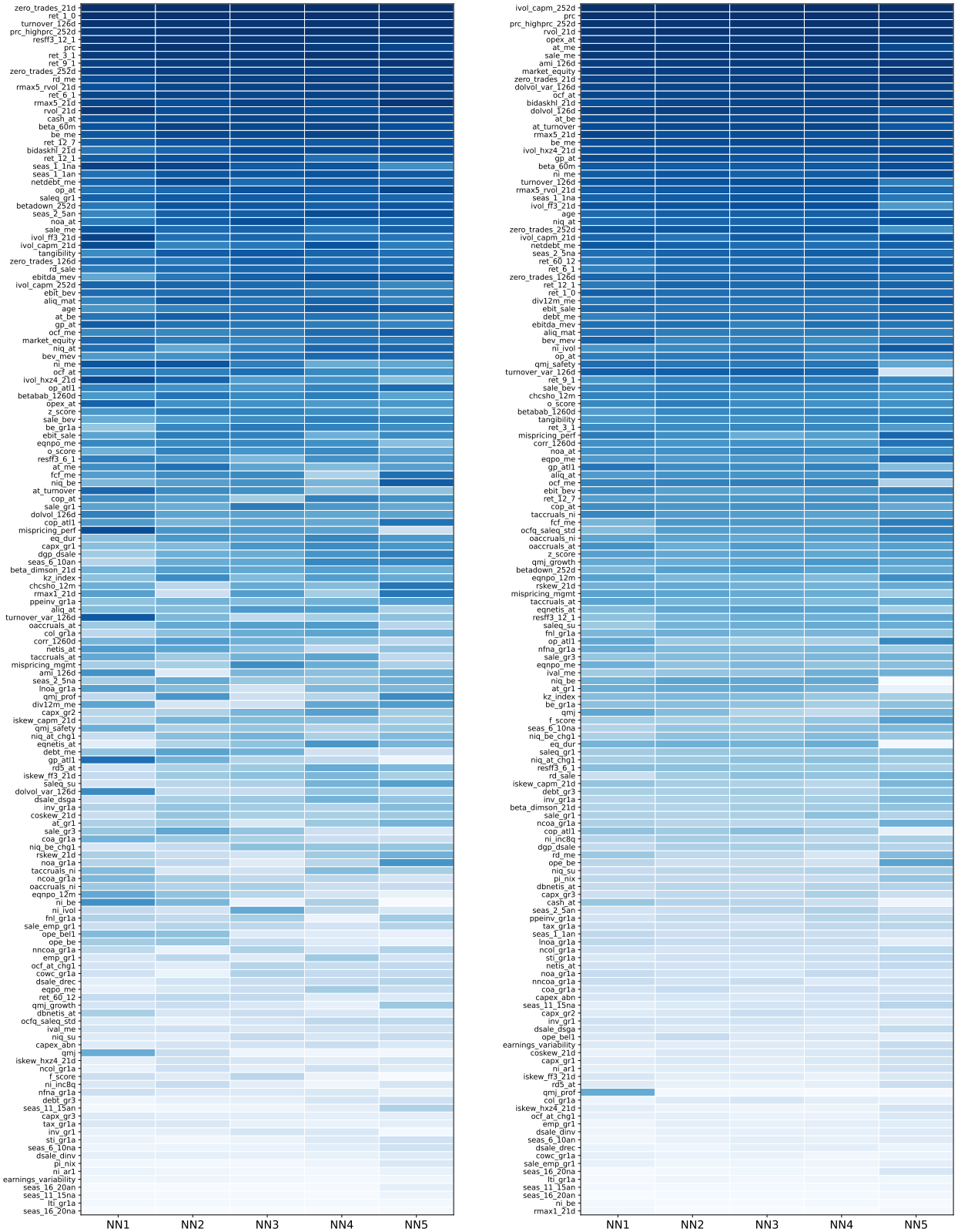
(b) For conditional volatility prediction

Figure 4: The top 20 variables with the greatest variable importance.

prediction, we first compute the risk premium predictions as usual. We then assess the VI by setting the values of the j-th variable to zero while holding other variables unchanged and inputting this configuration into the volatility component of the model. The VI in this context is quantified by the reduction in the negative log-likelihood loss.

Figure 4 presents the top 20 variables by importance for predicting risk premiums and conditional volatilities, respectively. The variable importance is averaged across NN1-–NN5. In both subfigures, the top two variables—zero_trades_21d and ret_1_0 for risk premium prediction, and ivol_capm_252d and prc for conditional volatility prediction—exhibit much greater importance compared to others. Figure 5 shows the VI rankings provided by all five network structures, with variables displayed in the order determined by the average ranking. In this figure, more important variables are positioned at the top, with darker colors indicating higher ranks.

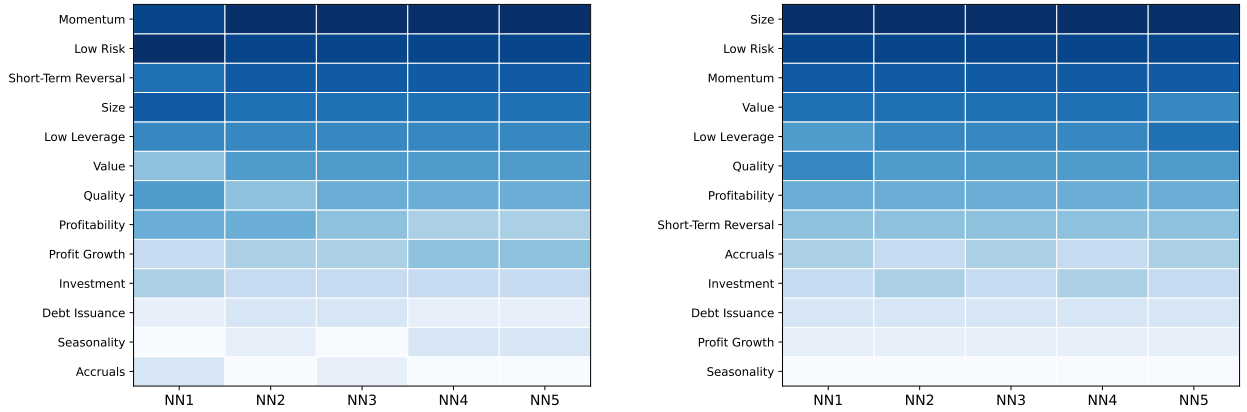
It is noteworthy that the importance ranking of all variables demonstrates a consistent pattern across different neural network structures. On the risk premium side, as illustrated in Figure 5a, the four most crucial variables are: number of zero trades with turnover as tiebreaker within 1 month (zero_trades_21d), short-term reversal (ret_1_0), share turnover (turnover_126d), and current price to high price over last year (prc_highprc_252d). We compare them with the results from Gu et al. (2020), Cakici et al. (2023), and Leippold et al. (2022), revealing some consistencies and differences. For example, both our study and Gu et al. (2020) examine the U.S. market, albeit over different time periods. The second most important variable in our study, ret_1_0, corresponds to the top variable identified by Gu et al. (2020). However, the two other significant variables from Gu et al. (2020), market_equity and ret_12_1, are ranked only in the mid to high range in our importance ranking (we rank ret_3_1 and ret_9_1 more highly). Cakici et al. (2023) investigated variable importance in international stock markets, highlighting variables such as prc_highprc_252d and ret_1_0 for predicting expected returns, which aligns with our findings on the importance of these variables. Additionally, Leippold et al. (2022) focused on the Chinese stock market, and we similarly find zero trades variables (zero_trades_21d and zero_trades_252d) to be of significant importance.



(a) For risk premium prediction

(b) For conditional volatility prediction

Figure 5: The comprehensive variable importance rankings for 153 variables provided by all five models. On the left, the ranking pertains to the risk premium prediction, while the right side focuses on the conditional volatility prediction. Variables deemed more critical are positioned higher in the figures and are depicted with more intense colors.



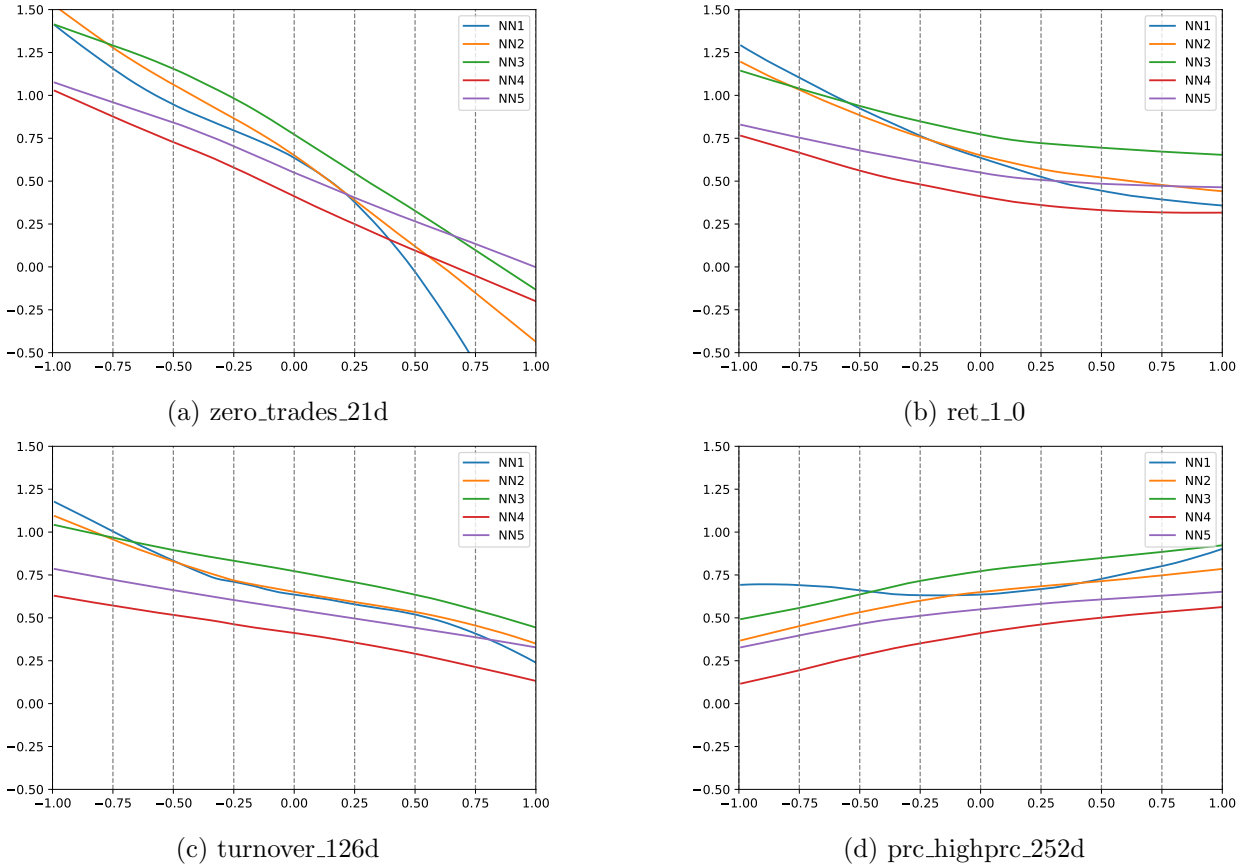
(a) For risk premium prediction

(b) For conditional volatility prediction

Figure 6: The comprehensive variable category importance rankings for 13 categories provided by all five models. On the left, the ranking pertains to the risk premium prediction, while the right side focuses on the conditional volatility prediction. Categories deemed more critical are positioned higher in the figures and are depicted with more intense colors.

On the conditional volatility side, as illustrated in Figure 5b, the four most important variables are: idiosyncratic volatility from the CAPM within 252 days (ivol_capm_252d), price per share (prc), current price to high price over last year (prc_highprc_252d), and return volatility (rvol_21d). Intuitively, volatility persistence is a well-established phenomenon, making historical volatility highly significant for predicting conditional volatility. Our analysis successfully identifies volatility-related variables, including ivol_capm_252d, rvol_21d, and the idiosyncratic volatility from the Hou–Xue–Zhang four-factor model (ivol_hxz4_21d) (Hou et al., 2015). Additionally, trading volume-related variables, such as dollar trading volume (dolvol_126d) and volatility of dollar trading volume (dolvol_var_126d), are also of considerable importance. This alignment with established intuitions about volatility prediction demonstrates that our method effectively captures key aspects of conditional volatility. Notably, several significant variables in this analysis, such as zero_trades_21d, prc_highprc_252d, and prc, also align with those identified as crucial for risk premium prediction.

In addition to analyzing the importance of individual variables, we investigate the importance of different variable categories. Using the categorization method proposed by Jensen et al. (2023), we group 153 individual variables into 13 categories based on similar economic attributes. We evaluate the importance of these variable categories separately for risk premium and conditional volatility predictions, by calculating the average importance of all individual variables within each category, as shown in Figure 6. The results in Figure 6a highlight the crucial role of the Momentum category in predicting risk premiums, followed by the Low Risk, Short-Term Reversal, Size, and Low Leverage categories. Gu et al. (2020) categorized their variables into four classes, with their most important category, recent price trends, aligning closely with our Momentum category, and their second most important category, liquidity, corresponding to our Low Risk category (many



(a) zero_trades_21d

(b) ret_1_0

(c) turnover_126d

(d) prc_highprc_252d

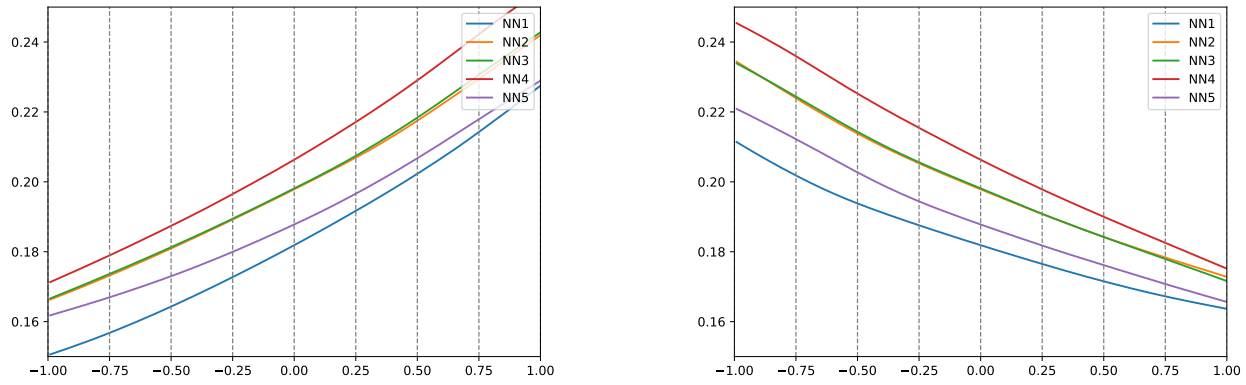
Figure 7: The marginal effects between characteristics and risk premium. The x-axis represents the values of a selected individual variable (from the four most important ones) ranging from -1 to 1, while the y-axis shows the corresponding predicted risk premium. All other characteristics are held constant at their median value of 0.

variables appear in both sets). Cakici et al. (2023), using the same categorization approach as ours, found that the most important category in the global markets is Value, with Momentum and Low Risk also being highly influential. For conditional volatility prediction, Figure 6b reveals that the most influential variable categories are Size, Low Risk, Momentum, and Value.

4.5.1. Marginal Effects

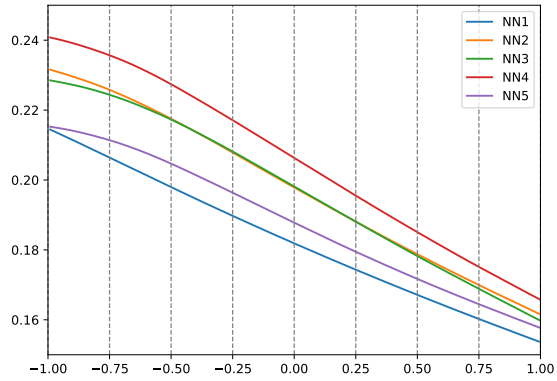
In this section, we examine the marginal effects between characteristics and risk premium, as well as between characteristics and conditional volatility. We focus on the four most important variables identified in Figure 5. For each selected variable, we vary its value within the range $[-1, 1]$, while keeping all other variables fixed at their median value of 0. We then analyze how these changes in the selected variable impact the predicted risk premium or conditional volatility.

Figure 7 illustrates the marginal effects between characteristics and risk premium. Overall, our findings are consistent with prior literature, with one specious exception. Specifically, a higher value of zero_trades_21d, indicating lower trading activity and thus reduced liquidity, is associated with a lower risk premium. This counterintuitive result may be due to the influence of other

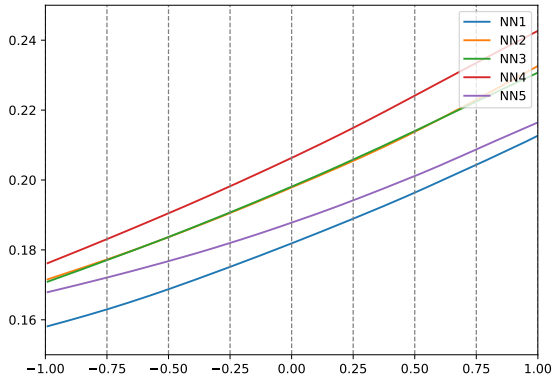


(a) ivol_capm_252d

(b) prc



(c) prc_highprc_252d



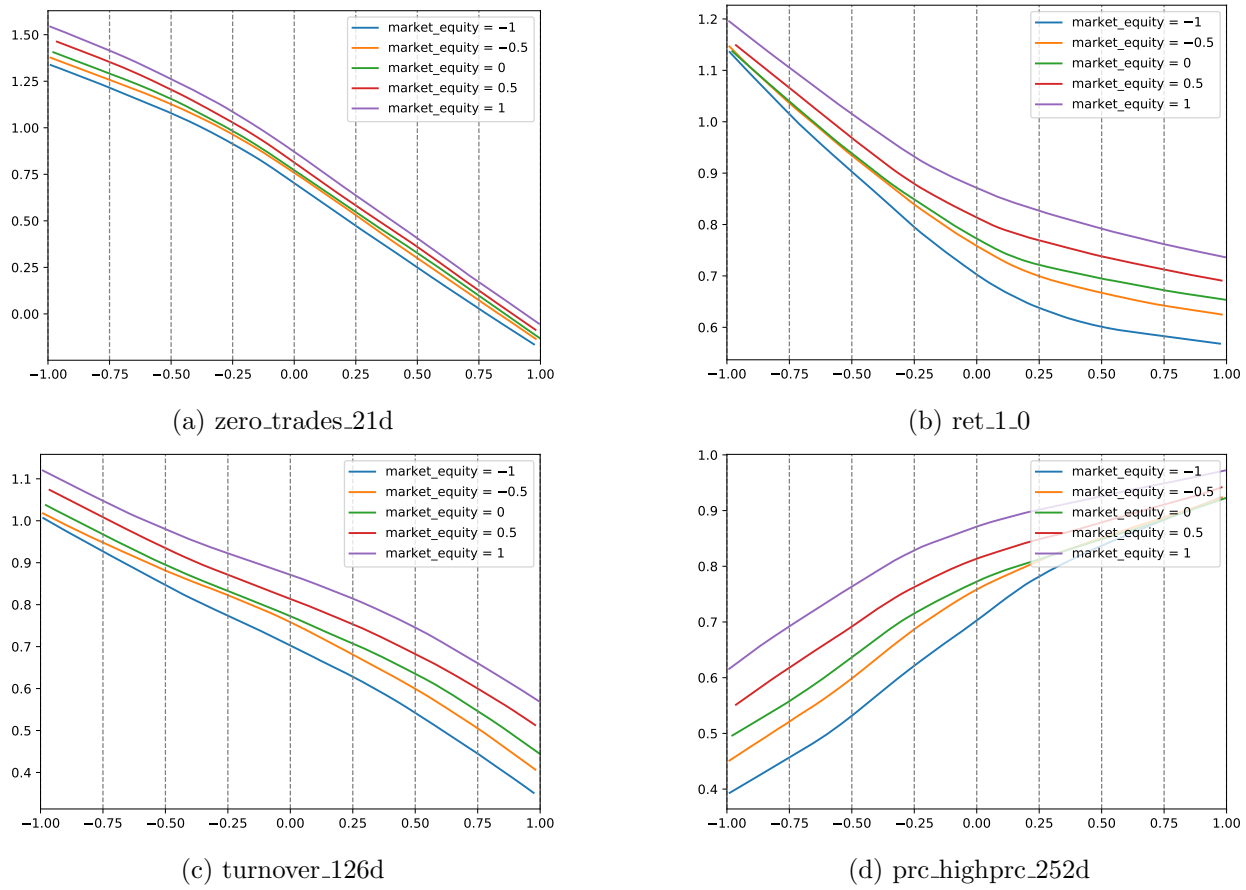
(d) rvol_21d

Figure 8: The marginal effects between characteristics and conditional volatility. The x-axis represents the values of a selected individual variable (from the four most important ones) ranging from -1 to 1, while the y-axis shows the corresponding predicted conditional volatility. All other characteristics are held constant at their median value of 0.

correlated variables. In contrast, the marginal effect of turnover_126d aligns with expectations, as share turnover is considered an alternative measure of liquidity. Additionally, our results indicate that stock returns decrease with the short-term reversal variable ret_1_0 and increase with the momentum variable prc_highprc_252d. These observations are consistent with the findings in Datar et al. (1998), Jegadeesh (1990), and George and Hwang (2004).

For the marginal effects between characteristics and conditional volatility, as demonstrated in Figure 8, we observe that conditional volatility is positively correlated with both idiosyncratic volatility (ivol_capm_252d) and return volatility (rvol_21d). This finding supports the effectiveness of our model. In contrast, there is a negative relationship between stock prices (prc or prc_highprc_252d) and conditional volatility, which is consistent with the well-established asymmetric volatility phenomenon.

In summary, Figures 7 and 8 demonstrate that different characteristics have distinct predictive relationships with risk premium and conditional volatility which align with prevailing financial theories and empirical findings. Additionally, these relationships are non-linear, and neural network



(a) zero_trades_21d

(b) ret_1_0

(c) turnover_126d

(d) prc_highprc_252d

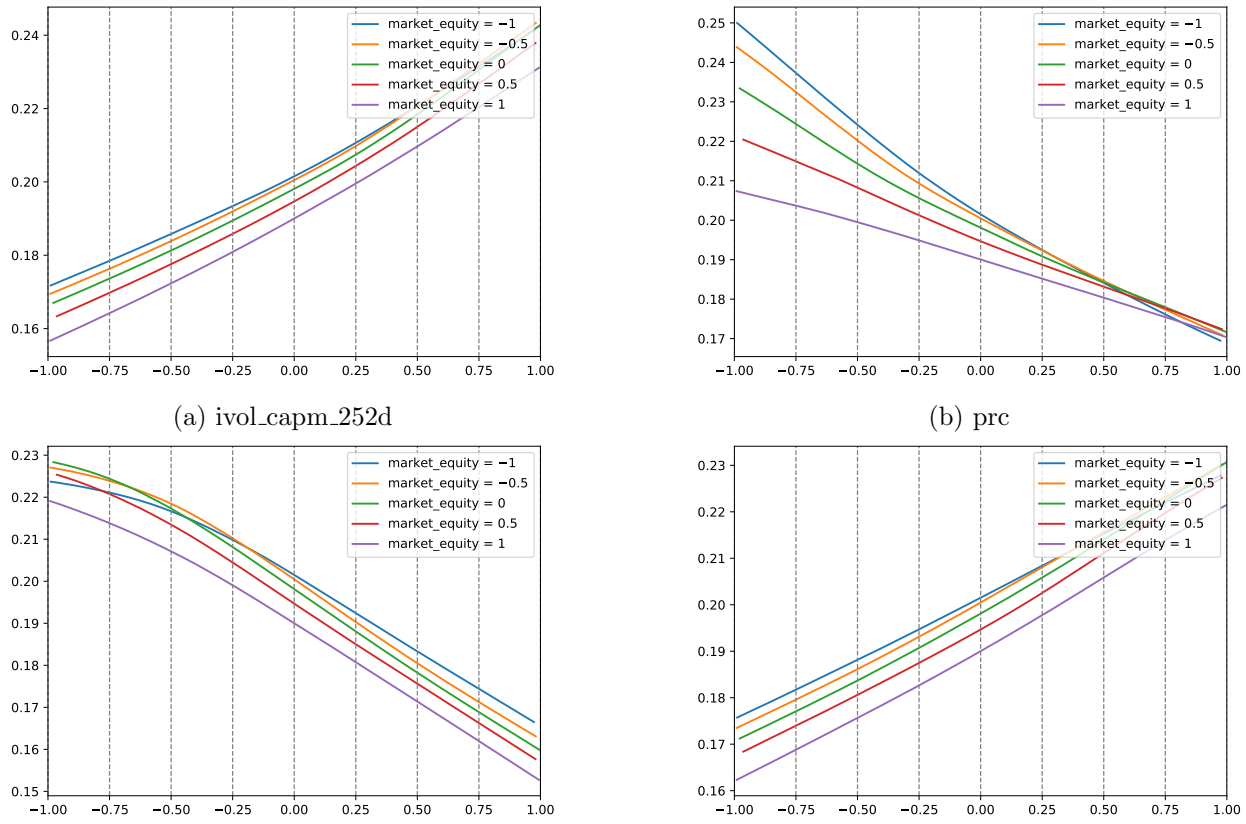
Figure 9: The interaction effects between market_equity and the selected variables for risk premium prediction. The x-axis represents the values of a selected individual variable (from the four most important ones) ranging from -1 to 1, while the y-axis shows the corresponding predicted risk premium when market_equity is set to a specific value. All other characteristics are held constant at their median value of 0. The results are averaged across all five neural networks.

models are capable of identifying such non-linear associations.

4.5.2. Interaction Effects

The functional form of the neural network model exhibits diversity, allowing it to capture not only the non-linear relationships between input variables and the target but also the interaction effects among variables. To analyze the interaction effects, we investigate how the marginal effects of a selected variable on risk premium and conditional volatility vary as we set stock market equity (market_equity) to different values. The selected variable is one of the four most important variables identified in Figure 5.

From Figure 9, we can gather information about the interactive effects among variables for risk premium prediction. Figure 9a suggests a lack of interaction between zero_trades_21d and market equity; Figure 9b demonstrates that the short-term reversal effect is more prominent in small-cap stocks (blue line), presenting a nonlinear marginal effect and a complex interaction; Figure



(a) ivol_capm_252d

(b) prc

(c) prc_highprc_252d

(d) rvol_21d

Figure 10: The interaction effects between market_equity and the selected variables for conditional volatility prediction. The x-axis represents the values of a selected individual variable (from the four most important ones) ranging from -1 to 1, while the y-axis shows the corresponding predicted conditional volatility when market_equity is set to a specific value. All other characteristics are held constant at their median value of 0. The results are averaged across all five neural networks.

9c indicates that share turnover exhibits similar but parallel relationships across different market equity values; and Figure 9d illustrates that the effect of prc_highprc_252d is more apparent in small-cap stocks, similar to the short-term reversal effect.

Figure 10 provides insights into the interactive effects among variables for conditional volatility prediction. Figure 10a presents five nearly parallel and closely spaced curves, suggesting relatively less interaction between market equity and idiosyncratic volatility in predicting conditional volatility; Figure 10b indicates that the price effects are more significant in small-cap stocks and this finding holds true only when the price is not high; and Figures 10c and 10d again present nearly parallel and closely spaced curves, indicating modest interaction effects.

5. Conclusion

In this paper, we introduce a new deep learning model that simultaneously forecasts stocks' risk premiums and conditional volatilities. From a predictive perspective, our model demonstrates

a strong ability to capture variations in both risk premiums and conditional volatilities, leading to promising forecasting results. This dual capability enhances our understanding of cross-sectional stock returns, enabling us to better manage risks in long-short portfolios and to gain insights into risk-return relationships.

We utilize our model's predictions to construct zero-investment long-short portfolios using double sorting methods. When compared to portfolios constructed through single sorting based solely on predicted risk premiums, our approach yields lower volatilities and significantly higher Sharpe ratios. Specifically, under the equal-weighted scheme, the maximum Sharpe ratio reaches approximately 3, while under the value-weighted scheme, it attains about 1.5. The risk-adjusted performance of these portfolios is also examined, revealing significantly positive alphas, which indicate that the superior performance cannot be explained by existing factors. Regarding the risk-return trade-off, we investigate the correlations between risk premiums and conditional volatilities in cross-sections based on our predictions. Our findings uncover a statistically significant and persistent negative correlation between them, offering a potential explanation for the superior performance of the zero-investment long-short portfolios constructed using our approach. Specifically, the volatilities of the long positions and of the short positions are effectively balanced, a contrast to what is observed with single sorting.

For the importance of characteristics, our analysis reveals that the risk premium prediction component of the model is significantly influenced by variables associated with Momentum, Low Risk, and Short-Term Reversal categories. In addition, the conditional volatility prediction component is predominantly driven by variables related to Size, Low Risk, and Momentum categories. These empirical findings align with previous studies that use machine learning to analyze stock markets. Furthermore, the marginal and interaction effects of the most important variables show strong consistency with established empirical phenomena, such as short-term reversal, momentum, volatility persistence, and asymmetric volatility. Overall, the results in this article highlight the effectiveness of deep neural network models in addressing empirical asset pricing challenges and contribute new insights to the evolving field of financial technology.

References

Aït-Sahalia, Y., Fan, J., Xue, L., Zhou, Y., 2022. How and When are High-Frequency Stock Returns Predictable? Technical Report. National Bureau of Economic Research.

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2006. The cross-section of volatility and expected returns. The Journal of Finance 61, 259–299.

Ang, A., Hodrick, R.J., Xing, Y., Zhang, X., 2009. High idiosyncratic volatility and low returns: International and further us evidence. Journal of Financial Economics 91, 1–23.

Bali, T.G., Cakici, N., 2008. Idiosyncratic volatility and the cross section of expected returns. Journal of Financial and Quantitative Analysis 43, 29–58.

Bali, T.G., Engle, R.F., Murray, S., 2016. Empirical asset pricing: The cross section of stock returns. John Wiley & Sons.

Bali, T.G., Goyal, A., Huang, D., Jiang, F., Wen, Q., 2020. Predicting corporate bond returns: Merton meets machine learning. Georgetown McDonough School of Business Research Paper , 20–110.

Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. The Review of Financial Studies 34, 1046–1089.

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics 31, 307–327.

Bollerslev, T., Hood, B., Huss, J., Pedersen, L.H., 2018. Risk everywhere: Modeling and managing volatility. The Review of Financial Studies 31, 2729–2773.

Bollerslev, T., Patton, A.J., Quaedvlieg, R., 2016. Exploiting the errors: A simple approach for improved volatility forecasting. Journal of Econometrics 192, 1–18.

Brandt, M.W., Kang, Q., 2004. On the relationship between the conditional mean and volatility of stock returns: A latent var approach. Journal of Financial Economics 72, 217–257.

Bryzgalova, S., Pelger, M., Zhu, J., 2020. Forest through the trees: Building cross-sections of stock returns. Available at SSRN 3493458 .

Cakici, N., Fieberg, C., Metko, D., Zaremba, A., 2023. Machine learning goes global: Cross-sectional return predictability in international stock markets. Journal of Economic Dynamics and Control 155, 104725.

Carr, P., Wu, L., Zhang, Z., 2020. Using machine learning to predict realized variance. Journal of Investment Management 18, 57–72.

Chen, L., Pelger, M., Zhu, J., 2024. Deep learning in asset pricing. Management Science 70, 714–750.

Choi, D., Jiang, W., Zhang, C., 2023. Alpha go everywhere: Machine learning and international stock returns. Available at SSRN 3489679 .

Cong, L.W., Feng, G., He, J., He, X., 2023. Asset pricing with panel tree under global split criteria. Technical Report. National Bureau of Economic Research.

Corsi, F., 2009. A simple approximate long-memory model of realized volatility. Journal of Financial Econometrics 7, 174–196.

Datar, V.T., Naik, N.Y., Radcliffe, R., 1998. Liquidity and stock returns: An alternative test. Journal of Financial Markets 1, 203–219.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica: Journal of the Econometric Society , 987–1007.

Engle, R.F., Ghysels, E., Sohn, B., 2013. Stock market volatility and macroeconomic fundamentals. Review of Economics and Statistics 95, 776–797.

Engle, R.F., Lilien, D.M., Robins, R.P., 1987. Estimating time varying risk premia in the term structure: The arch-m model. Econometrica: Journal of the Econometric Society , 391–407.

Engle, R.F., Mustafa, C., 1992. Implied arch models from options prices. Journal of Econometrics 52, 289–311.

Engle, R.F., Ng, V.K., 1993. Time-varying volatility and the dynamic behavior of the term structure. Journal of Money, Credit, and Banking 25, 336.

Fama, E.F., French, K.R., 2015. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. The Review of Financial Studies 33, 2326–2377.

Fu, F., 2009. Idiosyncratic risk and the cross-section of expected stock returns. Journal of Financial Economics 91, 24–37.

George, T.J., Hwang, C.Y., 2004. The 52-week high and momentum investing. The Journal of Finance 59, 2145–2176.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2006. Predicting volatility: getting the most out of return data sampled at different frequencies. Journal of Econometrics 131, 59–95.

Glosten, L.R., Jagannathan, R., Runkle, D.E., 1993. On the relation between the expected value and the volatility of the nominal excess return on stocks. The Journal of Finance 48, 1779–1801.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102, 359–378.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Advances in Neural Information Processing Systems 27.

Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.

Gu, S., Kelly, B., Xiu, D., 2021. Autoencoder asset pricing models. Journal of Econometrics 222, 429–450.

Guo, H., Savickas, R., 2010. Relation between time-series and cross-sectional effects of idiosyncratic variance on stock returns. Journal of Banking & Finance 34, 1637–1649.

Harvey, C.R., 2001. The specification of conditional expectations. Journal of Empirical Finance 8, 573–637.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780.

Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: An investment approach. The Review of Financial Studies 28, 650–705.

Jegadeesh, N., 1990. Evidence of predictable behavior of security returns. The Journal of Finance 45, 881–898.

Jensen, T.I., Kelly, B., Pedersen, L.H., 2023. Is there a replication crisis in finance? The Journal of Finance 78, 2465–2518.

Jiang, G.J., Xu, D., Yao, T., 2009. The information content of idiosyncratic volatility. Journal of Financial and Quantitative Analysis 44, 1–28.

Kelly, B., Xiu, D., et al., 2023. Financial machine learning. Foundations and Trends® in Finance 13, 205–363.

Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: A unified model of risk and return. Journal of Financial Economics 134, 501–524.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems 30.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.

Leippold, M., Wang, Q., Zhou, W., 2022. Machine learning in the chinese stock market. Journal of Financial Economics 145, 64–82.

León, A., Nave, J.M., Rubio, G., 2007. The relationship between risk and expected return in europe. Journal of Banking & Finance 31, 495–512.

Lochstoer, L.A., Muir, T., 2022. Volatility expectations and returns. The Journal of Finance 77, 1055–1096.

Ludvigson, S.C., Ng, S., 2007. The empirical risk–return relation: A factor analysis approach. Journal of Financial Economics 83, 171–222.

Luong, C., Dokuchaev, N., 2018. Forecasting of realised volatility with the random forests algorithm. Journal of Risk and Financial Management 11, 61.

Merton, R.C., 1973. An intertemporal capital asset pricing model. Econometrica: Journal of the Econometric Society , 867–887.

Merton, R.C., 1987. A simple model of capital market equilibrium with incomplete information. The Journal of Finance 42, 483–510.

Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. Econometrica: Journal of the Econometric Society , 347–370.

Nix, D.A., Weigend, A.S., 1994. Estimating the mean and variance of the target probability distribution, in: Proceedings of IEEE International Conference on Neural Networks (ICNN'94), IEEE. pp. 55–60.

Pagan, A.R., Schwert, G.W., 1990. Alternative models for conditional stock volatility. Journal of Econometrics 45, 267–290.

Pástor, L., Sinha, M., Swaminathan, B., 2008. Estimating the intertemporal risk–return tradeoff using the implied cost of capital. The Journal of Finance 63, 2859–2897.

Patton, A.J., Sheppard, K., 2015. Good volatility, bad volatility: Signed jumps and the persistence of volatility. Review of Economics and Statistics 97, 683–697.

Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. The Review of Financial Studies 23, 821–862.

Shen, Z., Xiu, D., 2024. Can machines learn weak signals? University of Chicago, Becker Friedman Institute for Economics Working Paper .

Wu, Q., Yan, X., 2019. Capturing deep tail risk via sequential learning of quantile dynamics. Journal of Economic Dynamics and Control 109, 103771.