JOIM
www.joim.com

JOURNAL OF INVESTMENT MANAGEMENT, Vol. 24, No. 4, (2026), pp. 4–30
© JOIM 2026

DEEP LEARNING FROM IMPLIED VOLATILITY SURFACES

Bryan Kelly[a], *Boris Kuznetsov*[b], *Semyon Malamud*[c] *and Teng Andrea Xu*[d]

We develop a novel methodology for extracting information from option implied volatility (IV) surfaces for the cross-section of stock returns, using image recognition techniques from machine learning (ML). The predictive information we identify shows minimal correlation with the existing option-implied characteristics, delivers a higher Sharpe ratio, and has a significant alpha relative to a battery of standard and option-implied factors. Large ML ensembles yield the best results, with performance improving as ensemble size increases. We introduce principal linear features, *an analog of principal components for ML, and use them to show* IV feature complexity: *A low-rank rotation of the IV surface cannot explain the model performance. Our results are robust to short-sale constraints and transaction costs.*

1 Introduction

Option Implied Volatility (IV) surfaces contain information about state-contingent risk premia and the probability distribution of returns over multiple horizons. Stock and option traders systematically use IV shape, including its slope, skewness, and other geometric features, to infer market expectations and risk attitudes and make trading decisions. The option pricing theory provides a basis for such analysis: local properties of this surface indeed contain information about the underlying market dynamics. For example, the surface variation along the moneyness dimension can be used to uncover the Arrow-Debreu state prices Breeden and Litzenberger (1978a), while the surface slope along the maturity dimension contains information about volatility by the Dupire formula Davis (2011). In the language of machine learning, universal *local* features (nonlinear transformations that depend on neighboring strikes and maturities) of the IV surface can be used to extract useful information about the stochastic structure of returns.

[a]Bryan Kelly is at Yale School of Management, AQR Capital Management, and NBER.
www.bryankellyacademic.org.
[b]Boris Kuznetsov is at the Swiss Finance Institute, EPFL.
[c]Semyon Malamud is at the Swiss Finance Institute, EPFL, and CEPR and is a consultant to AQR.
[d]Teng Andrea Xu is at EPFL.

Despite the underlying theory's elegance, empirically computing the theory-driven features is often associated with insurmountable econometric difficulties: theoretical formulas rely heavily on the continuity of strike and maturity dimensions, while, in reality, both strikes and maturities are discrete, living on a sparse grid. Furthermore, high bid-ask spreads due to often extreme illiquidity of options markets[1] introduce large amounts of noise into the estimation of equity implied volatilities.

In this paper, we leverage the progress in deep learning for computer vision and image recognition to construct powerful, non-linear, local features of IV surfaces. To this end, we exploit a particular neural network architecture called convolutional neural networks (CNNs). Thanks to their use of convolutional layers, CNNs are designed to capture spatial patterns and relationships between pixels in images. These layers apply filters to local regions of the input image, allowing the network to learn and identify patterns at various levels of abstraction (e.g., edges, shapes, and more complex features). The behavior of CNNs closely resembles how humans interpret images. As a result, CNNs have the potential to detect patterns on the IV surface that are similar to those recognized by professional traders during visual inspection. Furthermore, the locality of CNNs makes them perfect instruments for learning non-linear filters, extracting the volatility and risk premia information from the noisy IV surface[2].

We train several CNN architectures of increasing depth (complexity) on the standard OptionMetrics dataset of IV surfaces for several thousand stocks to predict stock returns over a one-month horizon. Since CNNs (like any neural networks) are trained by gradient descent, they are sensitive to (random) weight initialization: Depending on where the gradient descent starts, it may converge to a weight vector corresponding to a different local minimum[3]. A classic approach for dealing with this is to create an *ensemble* of neural networks corresponding to different weight initializations and then combine them, see Lakshminarayanan *et al.* (2017). We find that the benefits from this ensembling are huge because predictive models generated by different initializations produce portfolios with low pairwise correlations. As a result, for the most complex (4- and 5-layer) deep learning architectures, increasing the ensemble size from one to a hundred leads to an increase in the out-of-sample Sharpe ratio from 0.9 to 2.7 for the full stock universe. By contrast, for the lower complexity (1 hidden layer) CNN, the Sharpe ratio only increases from 0.80 to 1.6. As Lakshminarayanan *et al.* (2017) explain, the sensitivity of predictions generated by NN to initialization captures the degree of uncertainty around these predictions. In low signal-to-noise ratio environments of financial markets, this uncertainty is very high, implying large gains from ensembling.

The complexity of our CNN-based model seems to contradict the conventional principle of parsimony in economics, suggesting that predictive information (originating, e.g., from risk premia or behavioral anomalies) can be encoded into a small number of factors and stock characteristics. When applied to the IV surface, the principle of parsimony suggests that only a few key surface features might contain useful information about future stock returns. Most of these features used in the existing literature are constructed as simple, linear combinations of implied volatilities, for example, the IV level, the slope of the term structure, and the smile. To test whether such a low-dimensional structure is indeed present in our highly non-linear predictive model, we use the methodology of Constantine *et al.* (2014) to identify key linear combinations of IVs with

the most explanatory power. These combinations are natural analogs of principal components for our highly non-linear models, and we refer to them as *principal linear features*. Contrary to conventional wisdom, we find no evidence for a low-dimensional structure. While broadly consistent with the complexity principle of Kelly *et al.* (2021) and Didisheim *et al.* (2023), the nature of this phenomenon is different. Indeed, the results in Kelly *et al.* (2021) and Didisheim *et al.* (2023) mostly concern complexity due to non-linearities, whereas our last finding is about *feature complexity,* expressing the fact that a vast number of linear features is necessary to extract the predictive information contained in the IV surface.

As Kelly *et al.* (2021) and Didisheim *et al.* (2023) explain, *virtue of complexity* means that, even for an already complex model, we can often find a new, nonlinear transformation of the IV surface that boosts the out-of-sample performance. This complexity is not a puzzle to be solved or evidence of data mining. Instead, it is the theoretically expected outcome of learning a non-linear, high-dimensional relationship with limited data. The only constraint we impose on the model is the principle of locality, formalized by the choice of a convolutional NN instead of a generic NN architecture utilized, say, in Gu *et al.* (2020) and Chen *et al.* (2021).

Recently, several papers, e.g., DeMiguel *et al.* (2020), Detzel *et al.* (2023), and Muravyev *et al.* (2022), have argued that the performance of modern multi-factor models needs to be evaluated by accounting for transaction costs and short-sale constraints. We perform a detailed analysis of trading costs for our strategies and find that, perhaps surprisingly, the market-neutral portfolio constructed by sorting stocks based on the CNN-predicted return makes most of the money with the long leg, implying that our findings are not sensitive to short-selling costs and constraints. Transaction costs incurred by our strategy are significant: As with many machine learning models, it exhibits a high turnover (about 80% per month). Following Jensen *et al.* (2022b) and Didisheim *et al.* (2023), we study the model performance separately for several groups of stocks, created based on their market capitalization (mega, large, small, micro, and the non-micro group constructed as the set of all stocks excluding the micro-cap group). While the micro-group dominates the performance, the non-micro group also delivers a very strong Sharpe ratio and a significant alpha. Both micro- and non-micro groups retain their alpha significance after accounting for realistic costs. It is also important to note that, in any case, the optionable stocks (i.e., stocks with options traded on them) are typically large and have a significant trading volume, implying that even the predictability we identify for micro stocks can be exploited with some meaningful arbitrage capital[4].

The rest of this paper is organized as follows. Section 2 reviews the related literature. Section 3 describes data and methodology. Section 4 provides the necessary background on CNNs. Our main empirical results are reported in Section 5. Section 6 investigates feature importance, introduces *principal linear features* and documents *the virtue of feature complexity*. Section 7 concludes.

2 Related Literature

Understanding why average returns differ across assets is a central question in finance. Over the last few decades, the search for stock characteristics that predict returns has led to the emergence of the (constantly growing) “factor zoo": A huge number of characteristics that contain information about the cross-section, see Cochrane (2011), Harvey *et al.* (2016), McLean and Pontiff (2016),

Hou *et al.* (2020), Feng *et al.* (2020), Jensen *et al.* (2022b), and Giglio *et al.* (2022) for a recent overview.

Many papers in this literature focus on predicting asset returns using complex, non-linear models; see Moritz and Zimmermann (2016), Chinco *et al.* (2019), Han *et al.* (2019), Chen *et al.* (2019), Bryzgalova *et al.* (2020), Liu *et al.* (2020), Gu *et al.* (2020), Kozak *et al.* (2020), Freyberger *et al.* (2020), Avramov *et al.* (2021), Guijarro-Ordonez *et al.* (2021), Leippold *et al.* (2022), Kelly *et al.* (2021) and Didisheim *et al.* (2023). Given the ever-growing complexity of these models (both in terms of the number of characteristics and the degree of non-linearity of the predictive relationships), several papers develop techniques to "shrink" the cross-section and find a sparse representation of the expected returns, either through a form of dimensionality reduction, e.g., by exploiting principal components as in Kelly *et al.* (2020), Kozak *et al.* (2018), Kozak *et al.* (2020), Lettau and Pelger (2020) and Giglio and Xiu (2021), or by imposing sparsity directly in the space of characteristics (see, e.g., Gu *et al.* (2020), Freyberger *et al.* (2020) and Bryzgalova *et al.* (2023)). While the evidence is mixed, recent findings of Didisheim *et al.* (2023) suggest that complexity is there to stay, and there might be no feasible way to find a sparse representation of expected returns. Namely, as Didisheim *et al.* (2023) show, the factor zoo is simply a statistical phenomenon originating from the small data problem: We do not have enough data to find the right low-dimensional representation of expected returns, even if it exists; hence, our best bet is to build the most complex model without imposing a sparse prior[5].

Most of the above-mentioned papers focus on stock characteristics that are either purely price-based (such as momentum, see Carhart *et al.* (1997)) or depend on company fundamentals such as the book-to-market ratio. By contrast, our paper focuses exclusively on the predictive information contained in the implied volatility surface, motivated by the idea that derivative prices provide an interesting lens to uncover rich information about the underlying assets and associated risks. The early contributions by Breeden and Litzenberger (1978b) and Banz and Miller (1978) show how Arrow-Debreu state prices can be recovered from the option prices. State prices contain information about risk premia and (subjective) physical probabilities as market participants anticipate. Under technical conditions, some information about these physical probabilities can be recovered; see, e.g., Ross (2015), Borovička *et al.* (2016), and Jensen *et al.* (2019). The idea of extracting useful forward-looking information from both individual equity and index options has been exploited in many papers. For example, Bali and Hovakimian (2009) show that the difference between implied (i.e., risk-neutral) and physical volatility, as well as the difference between the implied volatilities of near-the-money call and put options are both strong predictors of stock returns; Cremers and Weinbaum (2010) use deviations from put-call parity to predict stock returns; Johnson and So (2012) show that the ratio of the volume between options and stocks predicts stock returns at the one-week horizon; Chordia *et al.* (2020) reach a similar conclusion when examining the (signed) order flow in equity index options; An *et al.* (2014) find that stocks with large increases in call (put) implied volatilities over the previous month tend to have high (low) future returns; Andersen *et al.* (2015), Bollerslev and Todorov (2014); Bollerslev *et al.* (2015), Lin and Todorov (2019), Begin *et al.* (2020) and Han *et al.* (2020) show how the jump (tail) risk extracted from options prices predicts future stock returns; Baltussen *et al.* (2018) show that the volatility of implied volatility has predictive power for future stock returns, and Dew-Becker and Giglio (2020) show how to

estimate cross-sectional uncertainty from option prices.

Many option-implied predictors build on a potential risk-return relationship between the risk-neutral variance, variance risk-premium (VRP), and stock returns; see, e.g., Bollerslev *et al.* (2009), Feunou *et al.* (2018), Martin and Wagner (2019), Kilic and Shaliastovich (2019), Feunou *et al.* (2019), Tang (2019), Kadan and Tang (2020), Pederzoli (2020), and Duarte *et al.* (2022).

Several other papers propose characteristics of the IV surface that are related to the shape of the implied volatility smirk or corresponding risk-neutral skewness portfolios; see, e.g., Xing *et al.* (2010), Yan (2011), Conrad *et al.* (2013), Stilger *et al.* (2017), Jones *et al.* (2018), and Bali and Murray (2019), and Schneider *et al.* (2020). While many of these papers find that IV-surface-based skewness measures contain predictive information, the nature of this information seems extremely sensitive to precise details of factor construction. For example, Conrad *et al.* (2013) document a negative relation between ex-ante risk-neutral skewness and asset returns, while Stilger *et al.* (2017) finds a positive relation[6]. These findings suggest that the nature of predictive information might be extremely complex, making it difficult to pin down the precise underlying economic mechanism.

The paragraphs above suggest that, even in the smaller world of option-implied characteristics, we are facing the problem of an ever-growing factor zoo, with new predictive relationships constantly discovered by academic researchers. Two recent papers attempt to bring order into this option-implied factor zoo by combining the informational contents of the multiple factors identified in the existing literature. Bali *et al.* (2022) use five of those factors (based on the findings of Bali and Hovakimian (2009), Cremers and Weinbaum (2010), Xing *et al.* (2010), and An *et al.* (2014)). They provide evidence that a linear tangency portfolio of these factors is not spanned by standard stock characteristics-based managed portfolios. However, in a more recent paper, Neuhierl *et al.* (2022) provide evidence that the factors in Bali *et al.* (2022) are spanned by fundamental stock characteristics when more of those characteristics are included in the model. Neuhierl *et al.* (2022) then argue that machine learning methods can deal with the IV-surface-based factor zoo and extract useful information unspanned by standard factors. They consider 17 options-based characteristics, including five of Bali *et al.* (2022). Then, they apply the adaptive group LASSO methodology of Freyberger *et al.* (2020) to build non-linear predictive models based on these 17 characteristics. They provide strong evidence that only 4 out of 17 option characteristics contain information about future stock returns not spanned by a large set of more than 60 stock characteristics. These four economically important characteristics are all related to the shape of the IV smirk.

While Neuhierl *et al.* (2022) do develop a complex, non-linear, machine learning model, they utilize relatively small (only 17) ready, pre-built characteristics of the IV surface, motivated by sparsity considerations and the idea that a few linear features of the IV surface summarize its predictive information. In this paper, we follow a different approach. First, we completely abstract from the existing set of characteristics, such as smirk, skew, slope, etc.; instead, we take an agnostic approach and let the machine learning model decide which features of the IV surface are useful for predicting stock returns. Second, we do not try to impose any form of sparsity on the model. Instead, we build a model with an exorbitant number of features and parameters following the principle of the virtue of complexity, see Kelly *et al.* (2021) and Didisheim *et al.* (2023), combined with the ensembling. As explained above, both aspects are important: There is a large

amount of predictive information in the IV surface, and we need a large ensemble of models to leverage this information efficiently. Furthermore, our model exhibits a very high *feature complexity:* A small number of linear IV features cannot span our model's predictive content.

We complete this literature review by noting that implied volatility is closely related to the physical volatility (in fact, in the idealized Black-Scholes continuous time setting, the two should be identical) and, hence, any signal about the level of the IV surface is closely related to the physical idiosyncratic stock volatility. The latter contains predictive information about stock returns and belongs to the large family of "low-risk anomalies", see Ang *et al.* (2006), Harvey and Siddique (2000), Boyer *et al.* (2010), Frazzini and Pedersen (2014), Amaya *et al.* (2015) and Schneider *et al.* (2020). In particular, Schneider *et al.* (2020) argue that these low-risk anomalies reflect compensation for co-skewness risk. It is possible that the more complex predictive signals extracted from the IV surface by our model might also reflect a form of risk premium related to some jump, tail, or high-order moment risk. Understanding the connection of our non-linear signals with risk premia is an important direction for future research.

3 Data and Methodology

3.1 Data

Options data are from OptionMetrics IvyDB. The historical IvyDB Volatility_Surface dataset contains interpolated option implied volatility surfaces for a large set of firms for each trading day from January 1996 to December 2021. For interpolation, a proprietary kernel smoothing algorithm is applied by OptionMetrics across both

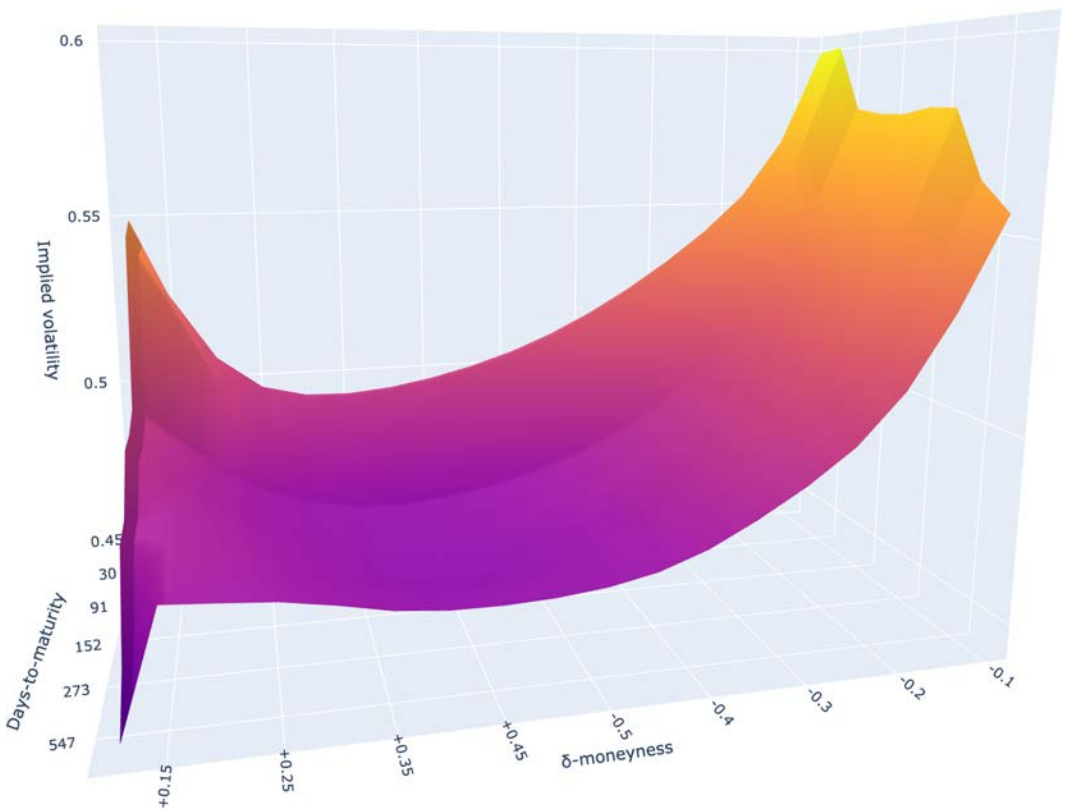


Figure 1 Transformed implied volatility surface of TSLA stock, 01/07/2016.

moneynesses (defined in terms of option δ) and expiry (defined in terms of days-to-maturity τ) grids, allowing us to abstract from option cycles and varying strikes. Thus, for each trading day, the δ-moneyness grid goes from -1 to $+1$ in equidistant 0.05 steps (puts have negative δ, calls have positive δ), and the expiry grid ranges from 10 to 730 days-to-maturity.

We hypothesize that in-the-money calls contain the least forward-looking information and only keep $\delta \in [-0.5, 0.5]$; then, we drop all 10-days-to-maturity values due to lack of data before 2010[7]; lastly, we re-stack volatility surfaces at-the-money (so that δ goes from 0.1 to 0.5, then goes from -0.5 to -0.1): this way, the combination of put and call IV components results in a more seamless transition. We end up with implied volatility surface images of size 10×18. Figure 1 shows a typical example of an implied volatility surface from Option Metrics IvyDB that we work with. We denote $IV_{i,t} = (IV_{i,t}(\delta, \tau))_{\delta\in\Delta,\tau\in\mathcal{T},} \in \mathbb{R}^{|\Delta|\times|\mathcal{T}|}$ the interpolated implied volatility surface of a stock $i \in N$ on the last day of the month $t \in T$. Here, $\delta \in \Delta$, where Δ is the grid of available δ-levels, and $\tau \in \mathcal{T}$, where $\mathcal{T}$ is the grid of available times to maturity.

We use daily stock data from CRSP to construct monthly total stock returns for all NYSE, AMEX, and NASDAQ firms. The CRSP sample is aligned with our OptionsMetrics sample and spans January 1996 to December 2021. The number of stocks with available stock and options data is shown in Figure 2. The total number of stocks in our sample is close to 25'000, with the average number of stocks at any point in time exceeding 3'000[8]. We denote $R_{i,t+1}$ the total return on stock i from the last business day of the month t to the last business day of the month $t+1$.

Monthly returns are constructed as follows. First, for each stock i and business day d, daily total gross returns are computed as $(1 + r_{i,d})$ or as $(1 + r_{i,d})(1+r_{i,d}^{\text{delist}})$ if the stock i delists on day d. Here, $r_{i,d}$ is the total daily return from the CRSP Daily Stock file, and $r_{i,d=delist}$ is the delisting return from the Delisting Information CRSP file. Missing daily returns[9] are replaced with zeros. Monthly total gross returns for a month t, denoted by $R_{i,t}$, are cumulative products of daily gross returns,

$$R_{i,t} = \prod_{d\in t}(1 + r_{i,d})(1 + r_{i,d}^{\text{delist}}\mathbf{1}_{d=\text{delist}}) - 1. \tag{1}$$

Everywhere in this paper, *we focus on predicting monthly returns* $R_{i,t+1}$ using only implied volatility surfaces *on the last business day of the preceding month* t, which we denote by $IV_{i,t}$[10].

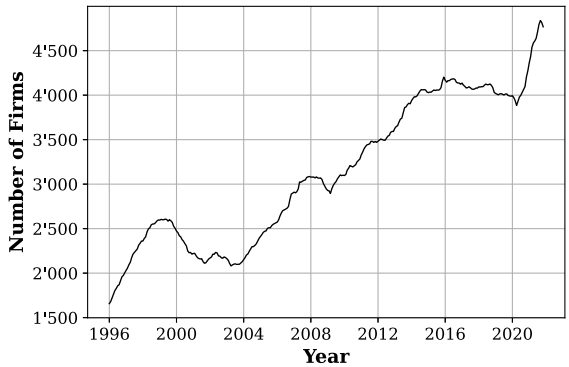


Figure 2 Evolution of the post-processed dataset size in terms of the number of firms.

3.2 *Ensembles of randomly initialized neural nets*

Given a family of non-linear functions, $f(x; w)$, indexed by a (high-dimensional) vector of parameters w (e.g., neural weights for the case of neural networks), we can optimize w to predict returns $R_{i,t+1}$ by plugging the month-end $IV_{i,t}$ into f and minimizing an error on the training data: Given a look-back horizon T, we can try selecting a parameter vector w that minimizes the in-sample prediction error based on the last T periods of the

data:

$$w_{*,t} = \arg\min_{w} \ell(w),$$

$$\ell(w) = \sum_{\theta=t-T+1}^{t} \sum_{i=1}^{N_\theta} (R_{i,\theta+1} - f(IV_{i,\theta}; w))^2, \tag{2}$$

where N_θ is the number of stocks in our data set at time θ. The standard way of finding w_* is by using gradient descent: Given a learning rate η, one can randomly initialize w_0 and then implement the algorithm of gradient descent by iteratively computing

$$\hat{w}_j(w_0) = \hat{w}_{j-1} - \eta \nabla_w \ell(\hat{w}_{j-1}) \tag{3}$$

for $j \geq 1$, where $\nabla_w \ell$ is the gradient of the loss $\ell(w)$ with respect to w.

By definition, the path of $\hat{w}_j(w_0)$ during the gradient descent (as j increases from one to ∞) depends on the initialization w_0. However, standard results imply that as the number j of gradient steps increases while the learning rate η converges to zero, $\hat{w}_j(w_0)$ converges to a *local extremum* $\hat{w}_\infty(w_0) = \lim_{j\to\infty,\eta\to 0} \hat{w}_j(w_0)$, satisfying $\nabla_w \ell(\hat{w}_\infty(w_0)) = 0$. As recent research shows, these local extrema are typically global minima for neural nets, see, e.g., Soudry and Carmon (2016). In simple, linear regression problems, the loss function $\ell(w)$ is convex in w and, hence, has a unique extremum which is also a global minimum; hence, for convex problems, $\hat{w}_\infty$ is independent of w_0. By contrast, for realistic, non-linear machine-learning models (such as the CNNs in our paper), $\ell(w)$ is not convex and has a tremendous number of (global or local) minima, see, e.g., Lakshminarayanan *et al.* (2017), Li *et al.* (2018), Yao *et al.* (2020) and Fort *et al.* (2019). As a result, $\hat{w}_\infty(w_0)$ depends on the initial weights, w_0, in a highly complex fashion. In particular, for K different initializations $w_0(k), k = 1, \ldots, K$, we will typically end up with K different weight vectors $\hat{w}_\infty(w_0(k))$. This leads to a *randomly initialized ensemble* of K models, $\{f(x; \hat{w}_\infty(w_0(k)))\}$, indexed by $k = 1, \ldots, K$. We then build the ensemble prediction by taking a simple average across k :

$$\widehat{R}^{\text{ens}}_{i,t} = \frac{1}{K} \sum_{k=1}^{K} f(IV_{i,t}; \hat{w}_\infty(w_0(k))). \tag{4}$$

Our findings present results for CNN ensembles of different sizes, with K ranging from 1 to 100. We observe a significant enhancement in portfolio performance as the ensemble size increases up to around 50 models, after which it saturates.

Given these predictions at each month-end, all stocks are sorted into deciles, equal-weighted portfolios to construct a long-short spread portfolio, "H-L", that is long the upper decile and short the lower decile. Denote $N_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}$ the number of stocks in the upper decile at time t, and $N_{\widehat{R}^{\text{ens}}_{i,t} < q_t}$ the number of stocks in the lower decile at time t, the returns on "H-L" are given by:

$$\begin{aligned} R^{H-L}_{t+1} &= \sum_{i=1}^{N_t} R_{i,t+1}\, w_{i,t}, \\ w_{i,t} &= \frac{\mathbf{1}_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}}{N_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}} - \frac{\mathbf{1}_{\widehat{R}^{\text{ens}}_{i,t} < q_t}}{N_{\widehat{R}^{\text{ens}}_{i,t} < q_t}}, \end{aligned} \tag{5}$$

where Q_t and q_t are the 90% and 10% percentiles of the distribution of $\widehat{R}_{i,t}$ at time t.

The algorithm described above depends on two key objects: The look-back window T in (2) used for training the model and the family of non-linear functions $f(x; w)$. For the lookback window, we follow the approach of Gu *et al.* (2020) and use an expanding window (allowing us to use all available data at time t)[11].

We consider three different families of functions $f_\ell(x; w)$, $\ell \in \{CNN1, CNN4, CNN5\}$ in our

Table 1 The number of parameters for a single model, 40-ensemble, and 100-ensemble configurations for our set of CNN models. Similarly, 'Single c', '40-ensemble c', and '100-ensemble c' represent the total model complexity as defined in Kelly *et al.* (2021): $c := P/T$, where $T = 973'947$ denotes the total number of observations at the end of the sample period (all models are trained using an expanding window).

Model	Single	40-Ensemble	100-Ensemble	Single c	40-Ensemble c	100-Ensemble c
CNN1	1'921	76'840	192'100	0.002	0.079	0.197
CNN4	98'241	3'929'640	9'824'100	0.101	4.035	10.087
CNN5	394'561	15'782'440	39'456'100	0.405	16.205	40.512
NN1	23'809	952'360	2'380'900	0.024	0.978	2.445

analysis. Each family is a convolutional neural network (CNN, see Section 4 for details), with $CNNi$ equipped with i hidden convolutional layers, $i = 1, 4, 5$. The number of layers of the network defines its complexity, expressive power, and ability to approximate non-linear functions. The complexity of a CNN model (defined as the number of parameters divided by the sample size, see Kelly *et al.* (2021)) increases exponentially with depth, see Table 1 for details. Intuitively, we expect shallow CNNs to learn fewer and simpler features, whereas more complex CNNs will detect a greater variety of features. While such complex models are tremendously over-parametrized and severely overfit the data in-sample, their performance out-of-sample might be better due to the virtue of complexity, see Kelly *et al.* (2021), Didisheim *et al.* (2023), and Belkin (2021).

4 Convolutional Neural Networks

This section provides a concise overview of the rationale behind Convolutional Neural Networks (CNNs), the architecture of the proposed models, and the training methodology employed.[12] Additionally, we briefly discuss the discrete convolution operation and model optimization via stochastic gradient descent. Please refer to the Online Appendix for rigorous mathematical definitions.

4.1 CNN architecture

Convolutional Neural Networks (CNNs) continue to be the state-of-the-art approach for image classification, see Shankar *et al.* (2020). We outline some of the key advantages of utilizing CNNs.

(1) Spatial Hierarchy and Local Connectivity: CNNs are designed to capture spatial patterns and relationships between pixels in images, thanks to their use of convolutional layers. These layers apply filters to local regions of the input image, allowing the network to learn and identify patterns at various levels of abstraction (e.g., edges, shapes, and more complex features). In contrast, Deep Neural Networks (DNNs) treat all input features independently, losing the spatial information and relationships between neighboring pixels. This makes it difficult for DNNs to learn and recognize complex image patterns effectively.
(2) Translational and Rotational Invariance: CNNs inherently possess translation and rotational invariance, meaning that they can recognize patterns and features regardless of their position in the image. Conversely, as mentioned earlier, DNNs treat each input independently, resulting in the loss of information.

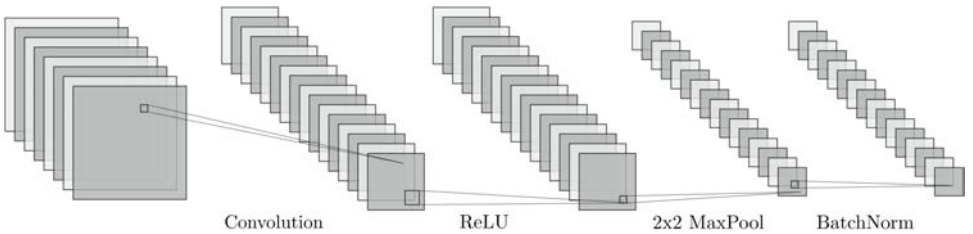


Figure 3 The figure above shows a building block of the CNN model consisting of a convolutional layer with a 3×3 filter, a ReLU layer, 2×2 max-pooling, and batch normalization layers. Note the max-pooling layer shrinks the height and width of the input by half and keeps the same depth.

We now describe a Convolutional Neural Network (CNN) as a sequence of operations to transform raw images, implied volatility surfaces in this case, into a prediction, see the Online Appendix for details.

A CNN core building block consists of three operations: convolution, activation, and pooling. In addition, we also use a batch normalization layer at the end of each block. A batch normalization layer normalizes the building core's output, reducing the so-called *internal covariance shift* (acting as a regularizer) from one building core to the next Ioffe and Szegedy (2015). The core CNN building block is shown in Fig. 3. A convolution layer applies filters to the input data, capturing local spatial patterns and generating feature maps as output. An activation function is a nonlinear function applied elementwise to the output of a layer. Max-pooling is a downsampling operation in a CNN that reduces the spatial dimensions of a feature map by selecting the maximum value within a defined neighborhood, helping to retain the most prominent features while reducing computational complexity and achieving translation invariance. Global Average Pooling is a pooling operation that computes the average value of each feature map across its entire spatial extent, effectively "vectorizing" the feature map. The final CNN layer is a single fully connected node that targets the next month's stock gross return. The CNNi architectures described above are constructed by stacking i such building blocks. For example, Figure 4 shows the CNN4 architecture.[13] See the Online Appendix for details and precise mathematical definitions of convolutions, MaxPool, and Global Average Pooling layers.

4.2 Training the CNN

Our initial training uses the first seven years of observations (1996–2002) in the expanding-window case. We call this the *warm-up period*. In this period, the CNN is trained for ten epochs. Then, we apply the trained model to the features of the subsequent month (unseen to the model) to make predictions. Afterwards, we include this new month in the training set to retrain the model using the same procedure. We call this the *transfer-learning period*. In this period, the CNN is trained for five epochs only, as the CNN has mostly learned the market dynamics during the *warm-up period* and only requires *fine-tuning* to adjust to an additional one month of observations.

We employ the same regularization methods as in Gu *et al.* (2020) to mitigate overfitting

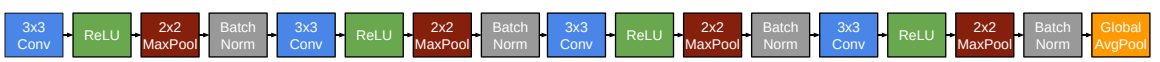


Figure 4 The figure above describes our CNN4 architecture.

and facilitate efficient computation. We use the Xavier initializer for weight initialization in each layer, as proposed by Glorot and Bengio (2010), which accelerates convergence by producing initial weight values that match the label scale. For loss function optimization, we combine stochastic gradient descent with the Adam algorithm Kingma and Ba (2014), setting the initial learning rate at 1×10^{-3} and using a batch size of 512.

5 CNN Portfolio Performance

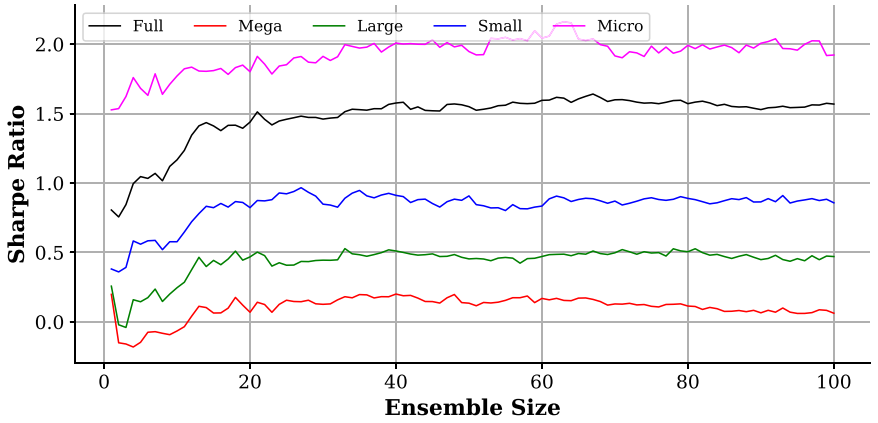
In this section, we investigate the out-of-sample performance of CNN-based return forecasts. Given a CNN architecture, we randomly initialize its weights w_0 using 100 different random seeds and train each of these NNs using gradient descent, as described in Section 4, thus creating an ensemble of CNNs. As described in Section 3.2, we construct a long-short spread portfolio, "H-L", that is long the upper decile and short the lower decile of predicted stock returns, with a monthly holding period, see (5). Then, we compute the predictions (4) gradually increasing the ensemble size K from 1 to 100, report the corresponding out-of-sample Sharpe ratios in Figure 5, and summarize their distribution in Table 2. Following Jensen *et al.* (2022b) and Didisheim *et al.* (2023), we perform the analysis separately for different size groups of stocks: mega (largest 20% of stocks based on NYSE monthly breakpoints), large (between 80% and 50% percentile of NYSE breakpoints), small (between 50% and 20% percentile of NYSE breakpoints), and micro (smallest 20% of stocks). Additionally, we report the results for the non-micro group that comprises all stocks with a market cap larger than the 20% NYSE percentile.

Figure 5 clearly illustrates that Sharpe ratios are monotone increasing in the ensemble size and, for more complex models (CNN4 and CNN5), saturate only around $K = 40$. While a single run of CNN1, CNN4, and CNN5 achieves an annualized Sharpe Ratio of 0.80, 1.48, and 0.88, respectively, the portfolios built using ensembles of $K = 100$ randomly initialized CNN1, CNN4, and CNN5 (see (4) for the definition), achieve out-of-sample Sharpe ratios of 1.64, 2.72, and 2.50, respectively. Figure 8 shows the cumulative performance of our strategy over time and compares it with the three strongest option characteristics-based factors from Neuhierl *et al.* (2022). As one can see, the complex model outperforms these factors.

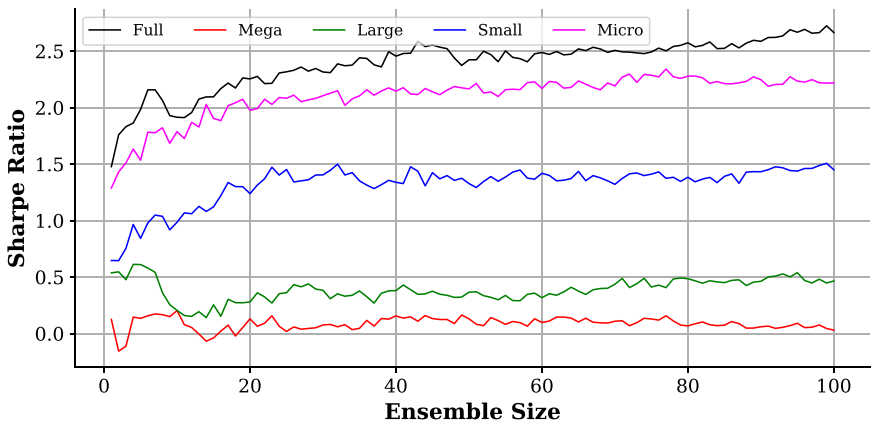
We now investigate to what extent standard and option-based characteristics span the performance of CNN-based factors. As Table 4 shows, ensemble CNN portfolios deliver a highly

Table 2 Minimum and maximum Sharpe Ratios for ensemble CNN1, CNN4, and CNN5 long-short portfolios from 2003/02 to 2022/12 from Figure 5.

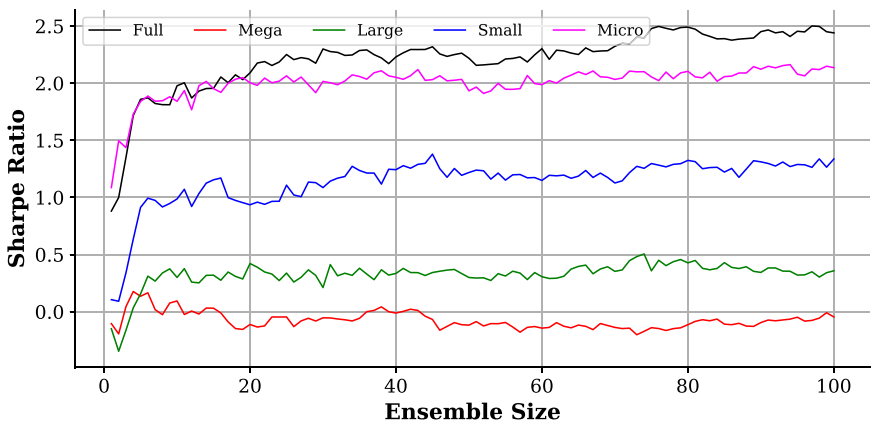
	CNN1		CNN4		CNN5	
	Min	Max	Min	Max	Min	Max
SR_{full}	0.756	1.641	1.477	2.724	0.881	2.501
SR_{mega}	−0.181	0.201	−0.156	0.204	−0.204	0.174
SR_{large}	−0.038	0.534	0.140	0.613	−0.340	0.511
SR_{small}	0.371	0.962	0.647	1.502	0.096	1.383
SR_{micro}	1.506	2.163	1.300	2.365	1.111	2.184
$SR_{\text{not_micro}}$	0.175	1.059	0.757	1.470	0.001	1.377



(a)



(b)



(c)

Figure 5 Sharpe ratios of ensemble-based returns (5) as a function of ensemble size. Figures 5(a), 5(b) and 5(c) show the performance of CNN1, CNN4, and CNN5, respectively.

Table 3 The table below shows our preprocessed Option Metrics statistics. Namely, Incomplete Images are those observations (stock and date pair) where some implied volatility value is missing, given a particular expiration date and delta pair. Missing Images is the sample where the stock and date pair appear in the dataset, but have all empty values. Complete Images are the correctly built images where the implied volatility surfaces are complete.

Year	Missing (%)	Incomplete (%)	Complete (%)	Total
1996	3289 (0.68%)	0 (0.0%)	477674 (99.32%)	480963
1997	5975 (1.03%)	0 (0.0%)	574945 (98.97%)	580920
1998	8789 (1.33%)	2 (0.0%)	651291 (98.67%)	660082
1999	11665 (1.7%)	10 (0.0%)	673334 (98.3%)	685009
2000	19943 (3.18%)	15 (0.0%)	606628 (96.81%)	626586
2001	28265 (4.84%)	6 (0.0%)	555407 (95.16%)	583678
2002	27345 (4.59%)	2 (0.0%)	568126 (95.41%)	595473
2003	10304 (1.84%)	5 (0.0%)	548341 (98.15%)	558650
2004	6137 (1.02%)	6 (0.0%)	593240 (98.98%)	599383
2005	6203 (0.94%)	3 (0.0%)	655095 (99.06%)	661301
2006	5991 (0.84%)	4 (0.0%)	706582 (99.16%)	712577
2007	8783 (1.12%)	5 (0.0%)	776060 (98.88%)	784848
2008	29436 (3.6%)	8 (0.0%)	788981 (96.4%)	818425
2009	40659 (5.04%)	7 (0.0%)	766591 (94.96%)	807257
2010	20822 (2.47%)	7 (0.0%)	821281 (97.53%)	842110
2011	26609 (2.94%)	9 (0.0%)	877955 (97.06%)	904573
2012	38756 (4.19%)	6 (0.0%)	886891 (95.81%)	925653
2013	31810 (3.22%)	3 (0.0%)	956879 (96.78%)	988692
2014	32621 (3.1%)	0 (0.0%)	1020567 (96.9%)	1053188
2015	51663 (4.78%)	0 (0.0%)	1030061 (95.22%)	1081724
2016	60616 (5.45%)	3 (0.0%)	1052121 (94.55%)	1112740
2017	53694 (4.95%)	1 (0.0%)	1031601 (95.05%)	1085296
2018	44705 (4.14%)	3 (0.0%)	1034656 (95.86%)	1079364
2019	54570 (5.1%)	1 (0.0%)	1016092 (94.9%)	1070663
2020	57409 (5.33%)	2 (0.0%)	1020423 (94.67%)	1077834
2021	20653 (1.63%)	0 (0.0%)	1248461 (98.37%)	1269114

significant alpha against a variety of option-based factor portfolios from Neuhierl *et al.* (2022) as well as the standard Fama-French factors. Complexity also has a very strong impact on the regression R^2 : While for the lower-complexity CNN1, the R^2 is about 66%, it drops to 36% for CNN4 and CNN5 models, suggesting that more complex models can pick up highly non-linear predictive patterns in the IV surface that are not accessible to simpler models.

Perhaps surprisingly, our most complex CNN4 and CNN5 models have only minimal and marginally significant exposure to option

Table 4 All stocks, long-short portfolio. Monthly OLS regression of the CNN1, CNN4, and CNN5 portfolios on the factor model that includes CIV, PIV, IVS_{atm}, IVS_{otm}, Skew, VOV, ΔCIV, and ΔPIV factor portfolios from Neuhierl *et al.* (2022), along with the standard Fama-French factors. The intercept coefficient is reported in monthly return terms, with corresponding standard errors in parentheses.

	CNN1	CNN4	CNN5
Intercept	0.008***	0.012***	0.011***
	(0.002)	(0.002)	(0.002)
$r_M - r_f$	−0.047	0.016	−0.048
	(0.061)	(0.056)	(0.055)
SMB	−0.183*	−0.040	0.021
	(0.100)	(0.090)	(0.089)
HML	−0.085	−0.082	−0.132*
	(0.085)	(0.077)	(0.076)
2-12 Momentum	−0.097**	−0.081*	−0.107**
	(0.046)	(0.042)	(0.041)
ST Reversal	−0.074	−0.137**	−0.175***
	(0.067)	(0.060)	(0.060)
LT Reversal	0.153*	0.173**	0.262***
	(0.087)	(0.078)	(0.077)
CIV	0.519**	0.120	0.053
	(0.231)	(0.209)	(0.206)
PIV	−0.304	−0.025	−0.014
	(0.236)	(0.213)	(0.210)
IVS_{atm}	−0.356***	0.088	0.245**
	(0.118)	(0.106)	(0.105)
IVS_{otm}	0.679***	0.181	0.170
	(0.133)	(0.121)	(0.119)
Skew	0.600***	0.244**	0.135
	(0.125)	(0.113)	(0.111)
VOV	0.159*	0.139*	0.144*
	(0.084)	(0.076)	(0.075)
Δ CIV	0.055	0.288***	0.182*
	(0.119)	(0.107)	(0.106)
Δ PIV	0.135	0.170*	0.086
	(0.108)	(0.098)	(0.096)
Observations	227	227	227
R^2	0.657	0.367	0.344
Adjusted R^2	0.634	0.325	0.301
Residual Std. Error	0.023	0.020	0.020
F Statistic	29.003***	8.766***	7.949***

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

characteristics-based factors. It has some exposure to long-term reversal while being significantly *negatively* exposed to short-term reversal, suggesting that the predictability we identify is not driven by standard short-term statistical arbitrage. Instead, exposure to long-term reversal suggests some link between our signals and fundamental stock valuations. The low-complexity CNN1 model exhibits significant exposure to the Skew factor, but as complexity increases, this exposure vanishes, confirming our intuition that complex models identify highly nonlinear patterns that are not spanned by standard linear characteristics of the IV surface.

The H-L strategy (5) is non-linear and exploits the power of the relationship between CNN-based predictions and future returns in the tails of the prediction distribution. We test the robustness of the above results by running simple, cross-sectional (Fama-MacBeth) regressions of returns on CNN-based predicted returns, using multiple characteristics as controls. Table OA.6 in the Online Appendix strongly supports our findings: Stock returns are indeed highly significantly related to CNN-based predictions, even after controlling for a large number of standard stock characteristics, including reversal, momentum, idiosyncratic volatility, as well as standard option-based characteristics.

We find that smaller-capitalization stocks largely drive the model performance, see Tables OA.1–OA.5 in the Online Appendix; nonetheless, CNN models still generate significant alpha for the not-micro group. Notably, the results from the Fama-Macbeth regressions per segment reveal that the statistical power of CNN forecasts is mainly concentrated in the small-cap stock group. This is highlighted in Tables OA.7 to OA.11. Lastly, value-weighted[14] portfolios built using CNN4 forecasts generate alphas that are statistically significant at a 10% level, see Table OA.33.

5.1 Simpler models

The above findings indicate that CNN-based models can extract important predictive information from the IV surface. But do we need CNNs? Is it possible to extract the same information using simpler models?

We start by investigating whether it is possible to use all IVs linearly to predict returns. To this end, we generate predictions based on the simple ridge *panel* regression model with an expanding window by building the prediction $\hat{R}_t^{\text{ridge}}$:

$$\hat{R}_t^{\text{ridge}} = IV_t(IV_{0:t-1}^{\top} IV_{0:t-1} + zI)^{-1} \times IV_{0:t-1}^{\top} R_{0:t-1}, \tag{6}$$

for each period in time $t = 0, \ldots, T$ with a ridge penalty z. Here, $IV_t \in \mathbb{R}^{N_t \times |\Delta| \times |\mathcal{T}|}$ represents the panel of implied volatility surfaces for N_t stocks at time t; $IV_{0:t-1} \in \mathbb{R}^{N_{0:t-1} \times |\Delta| \times |\mathcal{T}|}$ and $R_{0:t-1} \in \mathbb{R}^{N_{0:t-1}}$ denote the expanding window of implied volatility and true gross returns, respectively, covering the period $t = 0, \ldots, t-1$. This is a "kitchen sink" approach, whereby we are completely agnostic about the nature of predictors that the model generates. Table OA.12 reports the corresponding H-L strategy. As one can see, even such a simple strategy generates significant alpha relative to existing option-based factors. However, a per-group analysis, see Tables OA.13–OA.17, indicates that only the micro-group delivers significant alpha, in stark contrast to the case of CNN, see Tables OA.1–OA.5.

We now turn to a simpler non-linear model, referred to as NN1 in the sequel: a single-hidden-layer perceptron (a fully connected neural network). We use a relatively wide hidden layer of 128 neurons, implying a modest complexity (see Table 1). As for the CNNs, we build an ensemble of 100 randomly initialized NN1 models and find evidence for the same virtue of ensemble complexity as for the CNN models. Thus, we

build our final NN1 model by averaging across 100 seeds and then build the corresponding H-L portfolios. Table OA.18 studies whether the NN1-based factor absorbs any significant fraction of the alpha generated by the CNN models. Comparing with Table 4, we see that both alpha significance and magnitudes are preserved: While NN1 does absorb about 50% of the CNN1 alpha, it absorbs only 25% of the alpha of more complex CNN4 and CNN5 models. At the same time, Table OA.19 shows that the NN1 model does not exhibit any alpha relative to the simple CNN1 model. These findings have two important implications for us. First, they imply a very strong form of the virtue of complexity, manifesting itself in the ability of CNN4 and CNN5 models to identify non-linear features that the simpler models cannot capture. Second, ignoring the geometry of the IV surface and the key *locality principle* used by CNNs for feature construction leads to highly inferior performance of the simple NN1 model. Just naively building bigger models suggested by the complexity principle of Kelly *et al.* (2021) and Didisheim *et al.* (2023) is not enough: One needs to exploit the economic structure of the data and build neural architectures that optimally exploit this structure.

5.2 Long only

Short selling might be extremely costly for some stocks, especially less liquid ones. Many anomalies have been criticized for being difficult to implement for this reason, see Muravyev *et al.* (2022). In this section, we investigate the long leg of our strategy, defined as

$$R^H_{t+1} = \sum_{i=1}^{N_t} R_{i,t+1}\, w_{i,t}, \qquad w_{i,t} = \frac{\mathbf{1}_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}}{N_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}}, \tag{7}$$

where Q_t is the 90% percentile of the predicted returns, and $N_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}$ is the number of stocks in the upper decile at time t. The performance of this long-only strategy is reported in Figure 9 (cumulative returns plot) and Table 5 (regression statistics) for all stocks. To emphasize the power of our results, we set an extremely high bar for our long-only strategy and compute the alphas with respect to the benchmark factors (including all option-based factors) that are long-short. As one can see, the performance is robust, and alphas are highly significant. Compared to Table 4, complex models lose about half of their alpha and gain (not surprisingly) a huge exposure to the market portfolio.

5.3 Transaction and Short-Sale Costs

Many stock market anomalies and factors have been criticized for generating very high turnover, and, hence, their performance is extremely sensitive to transaction costs. Many characteristics-based portfolios generate negative performance after costs, see, for example, Detzel *et al.* (2023). As Jensen *et al.* (2022a) show, efficiently exploiting machine-learning-based strategies requires optimizing portfolio positions to reduce turnover optimally. While it is possible to use the methodology of Jensen *et al.* (2022a) and incorporate costs and turnover directly into our optimization algorithm, in this paper, we purposely follow a simpler approach and evaluate the performance of (5) and its long-only version (7) after costs, without applying position smoothing techniques.

We apply linear transaction fees[15] and short-sale costs to all our portfolios (CNN1, CNN4, CNN5, NN1, and Ridge) and, for a fair comparison, to all option-based factors from Neuhierl *et al.* (2022) against which we benchmark our models. We do not apply fees to the factors that are not option-based (e.g., momentum, reversal, Fama-French factors), which makes our analysis even

Table 5 All stocks, long-only portfolio (7). Monthly OLS regression of the CNN1, CNN4, and CNN5 long-only portfolios on the factor model that includes the CIV, PIV, IVS_{atm}, IVS_{otm}, Skew, VOV, ΔCIV, and ΔPIV factor portfolios from Neuhierl *et al.* (2022), along with the standard Fama-French factors. The intercept coefficient is reported in monthly return terms, with corresponding standard errors in parentheses.

	$CNN1_{ew}$	$CNN4_{ew}$	$CNN5_{ew}$
Intercept	0.004***	0.005***	0.005***
	(0.002)	(0.001)	(0.001)
$r_M - r_f$	0.835***	0.944***	0.930***
	(0.049)	(0.041)	(0.042)
SMB	0.248***	0.422***	0.473***
	(0.079)	(0.066)	(0.068)
HML	0.048	0.016	−0.014
	(0.067)	(0.056)	(0.058)
2-12 Momentum	−0.167***	−0.127***	−0.145***
	(0.037)	(0.031)	(0.032)
ST Reversal	0.048	0.015	0.008
	(0.053)	(0.044)	(0.046)
LT Reversal	0.099	0.130**	0.176***
	(0.068)	(0.057)	(0.059)
CIV	0.655***	0.568***	0.467***
	(0.183)	(0.153)	(0.158)
PIV	−0.105	−0.026	0.016
	(0.187)	(0.157)	(0.161)
IVS_{atm}	−0.299***	−0.157**	−0.050
	(0.093)	(0.078)	(0.080)
IVS_{otm}	0.401***	0.171*	0.142
	(0.105)	(0.088)	(0.091)
Skew	0.458***	0.279***	0.221**
	(0.099)	(0.083)	(0.085)
VOV	0.115*	0.125**	0.141**
	(0.066)	(0.056)	(0.057)
Δ CIV	0.035	0.112	0.029
	(0.094)	(0.079)	(0.081)
Δ PIV	−0.007	0.036	−0.022
	(0.085)	(0.072)	(0.074)
Observations	227	227	227
R^2	0.946	0.967	0.963
Adjusted R^2	0.943	0.965	0.961
Residual Std. Error	0.018	0.015	0.015
F Statistic	267.669***	444.960***	395.497***

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

more conservative. Following Crego *et al.* (2023), we apply the method that accounts for transaction costs at a portfolio level, which is described below. At a given time t, denote $w_t \in R^{N_t}$ the weight vector with coordinates

$$w_{i,t} = \frac{\mathbf{1}_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}}{N_{\widehat{R}^{\text{ens}}_{i,t} > Q_t}} - \frac{\mathbf{1}_{\widehat{R}^{\text{ens}}_{i,t} < q_t}}{N_{\widehat{R}^{\text{ens}}_{i,t} < q_t}}, \tag{8}$$

i.e., our portfolios (CNN1, CNN4, CNN5, NN1, and Ridge) are equal-weighted, and, at a given t, all positive (negative) weights sum up to 1 (-1). We abuse the notation and use $N_t \cup N_{t+1}$ to denote the set of stocks available for trading at times t and $t+1$. Then, the unit cost (per dollar of investment) of rebalancing and short-selling at time $t+1$ is given by

$$\begin{aligned}\bar{f}_{t+1} = &\sum_{i \in N_t \cup N_{t+1}} \left(f_i \left| w_{i,t+1} \right.\right. \\ &\left.\left. - \frac{1 + r_{i,t+1}}{1 + r_{t+1}} w_{i,t} \right| + (-w_{i,t})^+ \theta_i \right) \\ = \theta_i + &\sum_{i \in N_t \cup N_{t+1}} f_i \left| w_{i,t+1} \right. \\ &\left. - \frac{1 + r_{i,t+1}}{1 + r_{t+1}} w_{i,t} \right|, \end{aligned} \tag{9}$$

where r_{t+1} stands for the total return on the long-short portfolio with weights defined in (8) and $x^+ = \max(x, 0)$. Notice that we used that weights set at time t change from t to $t+1$ relative to the total portfolio value by $\frac{1+r_{i,t+1}}{1+r_{t+1}}$. This reflects the fact that the security value in the portfolio compounds from t to $t+1$ and the portfolio value compounds from t to $t+1$ as a whole. For simplicity, we assume that short-sale costs are paid at the beginning of each holding period (hence the use of $w_{i,t}$ in the term $(-w_{i,t})^+\theta_i$). We also used that $\sum_{i \in N_t \cup N_{t+1}} (-w_{i,t})^+ = 1$ by definition (8). Now, let I_t denote the total investment value of the portfolio at time t. The investment value evolves from $t+1$ to $t+2$ as follows:

$$I_{t+2} = I_{t+1}(1 - \bar{f}_{t+1}) r_{t+2}, \tag{10}$$

where r_{t+2} is a pre-fee gross return on the portfolio of stocks between $t+1$ and $t+2$, and I_{t+1} is pre-$(t+1)$-rebalancing investment value. Regrouping, we get the net-of-fee net return on the portfolio

$$\frac{I_{t+2}}{I_{t+1}} - 1 = (1 - \bar{f}_{t+1}) r_{t+2} - 1. \tag{11}$$

This simple identity allows us to deduct linear transaction costs without keeping track of the investment value, but by discounting net-of-fee returns only.

We set transaction costs f_i for a stock i in the micro-cap segment to be two times the cost level for other segments, so that

$$f_i = (1 + \mathbf{1}_{i \in \text{microcap}}) f_{\text{base}} \tag{12}$$

and we investigate two settings: $f_{\text{base}} = 0.001$ (ten basis points (bps)) and $f_{\text{base}} = 0.002$ (twenty bps). To account for costly short-selling, we assume a fixed monthly cost of

$$\theta_i = f_i. \tag{13}$$

For $f_{\text{base}} = 0.002$, this corresponds to 40 bps (= 4.8% per annum) short-sale cost for micro-stocks and 20 bps (=2.5% per annum) short-sale cost for non-micro stocks[16].

Again, to ensure a fair comparison, we apply the same fee structure to all option-based factor portfolios from Neuhierl *et al.* (2022), and we do not apply fees to other factors, which adds to the challenge for CNN-based portfolios to maintain their superiority.

Table OA.20 in the Online Appendix shows a per-segment breakdown of CNN, NN1, and ridge portfolios' Sharpe ratios without fees and at two different cost levels: 10 bps (and 20 bps for the micro-cap segment) and 20 bps (and 40 bps for the micro-cap segment). For both fee levels, the

micro-cap segment shows high and significant net-of-fee excess returns for CNNs, and, notably, CNN4 also exhibits consistent excess performance in the not-micro-cap segment at 5% (10%) significance level for the 10 (20) bps cost level, see Tables OA.21 to OA.32. On the other hand, the performance of NN1 and the ridge model diminishes, indicating that it's not solely the micro-cap segment that drives excess returns for CNNs. Overall, CNN4 withstands tests with conservative linear transaction and short-sale costs, even outside the micro-cap segment.

6 Principal Linear Features

Despite the abundance of data and the recent emergence of the "factor zoo" with numerous characteristics identified as return predictors, predictive relationships in economic and financial data are commonly believed to be sparse. While characteristic-based sparsity (the hypothesis that only a few characteristics matter for economic relationships) is likely an illusion Giannone *et al.* (2021), many papers argue that there exists some form of "*linear feature sparsity*", usually formulated in terms of the principal components Kozak *et al.* (2020). Namely, only a few top principal components are believed to be responsible for most of the predictable variation in returns.

Literature on the predictive content of the IV surface largely relies on the idea of linear feature sparsity and studies only a small number of signals represented as linear transformations of the IV surface, for instance, CIV, PIV, IVS_{atm}, IVS_{otm}, and related option characteristics described above are all examples of linear transformations of the IV surface. It is thus natural to ask a more general question of whether it is possible to use some non-parametric statistical techniques to extract relevant linear combinations of the IV surface with the most predictive content.[17] Under the linear feature sparsity hypothesis, we expect that the true dependence

$$f(IV_{i,t}) = E[R_{i,t+1}|IV_{i,t}] \tag{14}$$

of the expected returns on implied volatilities is given by a *low-rank function*, as formalized in the following definition.

Definition 1. Let $\mathcal{S} \subseteq \mathbb{R}^d$. A function $f : \mathcal{S} \to \mathbb{R}$ is said to have rank r on $\mathcal{S}$ if there exists a low-rank matrix $M \in \mathbb{R}^{r\times d}$, $r < d$, and a function $g : M\mathcal{S} \to \mathbb{R}$ such that $f(x) = g(Mx)\ \forall\, x \in \mathcal{S}$.

In the context of our machine learning task, $\hat{x} = Mx \in \mathbb{R}^r$ is the r-dimensional vector of *linear features*, i.e., those linear transformations of x that matter for the function f. To determine whether a given function is low-rank, we will use the following observation from Constantine *et al.* (2014) and Radhakrishnan *et al.* (2022):

Lemma 1. *Let $\mathcal{S} \subset \mathbb{R}^d$ be an open subset and $f(x) : \mathcal{S} \to \mathbb{R}$ be a real analytic function such that $\nabla f(x)$ is bounded on $\mathcal{S}$. Let also X be a random vector taking values in $\mathcal{S}$, such that its density $p(x) : \mathcal{S} \to \mathbb{R}_+$ is Lebesgue-almost surely positive on $\mathcal{S}$. Define*

$$M_* = E[\nabla f(X)\nabla f(X)'] \in \mathbb{R}^{d\times d}. \tag{15}$$

Suppose that rank $M_* = r$, *and let $M_* = UDU'$ be its eigenvalue decomposition, where $D = \text{diag}(\lambda_1, \ldots, \lambda_r)$ is the vector of non-zero eigenvalues of M_* and $U \in \mathbb{R}^{d\times r}$ is the matrix of corresponding eigenvectors. Then, for Lebesgue almost every x_0, $g(y) = f((I - UU')x_0 + Uy)$ satisfies $f(x) = g(U'x)\ \forall\, x \in \mathcal{S}$. Conversely, any function $f : \mathcal{S} \to \mathbb{R}$ of rank r has* rank $E[\nabla f(X)\nabla f(X)'] \le r$. *Furthermore, if $f(x) = g(Mx)$ where $M \in \mathbb{R}^{r\times d}$ has rank exactly r and $g(y)$ is real analytic on $M\mathcal{S}$ and the functions $\nabla_{y_i} g(y)$, $i = 1, \ldots, r$, are linearly independent, then,*

$$\text{rank } E[\nabla f(X)\nabla f(X)'] = r.$$

This lemma motivates the following definition:

Definition 2. We refer to the eigenvectors of the gradient outer product matrix (15) as the *principal linear features.*

By definition, principal linear features of f are the directions along which the function varies the most.

We now show how principal linear features can be used to capture a large fraction of the variation of the f function. To state the result, we will make use of the Poincaré inequality Leoni (2017): We say that a probability measure $p(x)dx$ on $\Omega \subset \mathbb{R}^d$ satisfies the Poincaré inequality on Ω if there exists a constant C such that

$$\text{Var}[f(x)] \leq C\, E[\|\nabla f(x)\|^2] \qquad (16)$$

for any continuously differentiable function f. Here, we use this inequality for Gaussian measure Brascamp and Lieb (1976): If $p(x)$ is a Gaussian measure with the covariance matrix Σ on $\Omega = \mathbb{R}^n$, then $C = \lambda_1(\Sigma)$, where λ_1 is the largest eigenvalue of Σ. It is possible to prove the following result using subtle properties of the Gaussian measures.

Algorithm 1. Linear Feature Sparsity

Require: Full sample of $IV_{i,t} \in \mathbb{R}^{18\times 10}$ for each stock $i \in N_t$ available at time t, and ensembles of CNNs $f_t^{\text{ens}}(\cdot)$ trained until t for all $t \in T$.

Compute the total number of $IV_{i,t}$ observations in the sample $N = \sum_t^T N_t$,
Compute average outer gradient product $M = \frac{1}{N}\sum_{i,t} \nabla f_T^{\text{ens}}(IV_{i,t}) \nabla f_T^{\text{ens}}(IV_{i,t})'$,
Perform eigenvalue decomposition of $M = UDU'$,
for $p \in [1, \ldots, 180]$ **do**
  Pick eigenvectors U_p corresponding to p largest eigenvalues, and denote U_{-p} the remainder,
  Let vec(x) be the vertical stack of flattened $IV_{i,t}$ for each i, t, i.e. vec$(x) \in \mathbb{R}^{N\times 180\times 1}$,
  Compute $\mu_y = \frac{1}{N}\sum_{i,t} U_p \text{vec}(x)$,
  Compute $\mu_z = \frac{1}{N}\sum_{i,t} U_{-p} \text{vec}(x)$,
  Compute $\Sigma_x = \text{Cov}(\text{vec}(x), \text{vec}(x)) \in \mathbb{R}^{180\times 180}$,
  Compute $\Sigma_z = U'_{-p}\Sigma_x U_{-p}$,
  Compute $\Sigma_{zy} = U'_{-p}\Sigma_x U_p$,
  Compute $\Sigma_y = U'_p \Sigma_x U_p$,
  Compute $\hat{\Sigma}_z = \Sigma_z - \Sigma_{zy}\Sigma_y^{-1}\Sigma_{yz}$,
  Sample $\epsilon \sim \mathcal{N}(0, 1) \in \mathbb{R}^{n\times(180-p)}$,
  for i, t **do**
    Compute vec$(y_{i,t}) = U'_p \text{vec}(x_{i,t})$, where
    Sample N_ε realisations of $\text{vec}(z|y_{i,t,k}) = \mu_z + \Sigma_{zy}\Sigma_y^{-1}(\text{vec}(y_{i,t}) - \mu_y) + \hat{\Sigma}_z^{1/2}\varepsilon_k$, $k = 1, \ldots, N_\varepsilon$,
    Estimate $f_{p,t}^{\text{ens}}(y_{i,t}) = \mathbb{E}[f_t^{\text{ens}}(z + U_p y_{i,t})|y_{i,t}] = \frac{1}{N_\varepsilon}\sum_{k=1}^{N_\varepsilon} f_t^{\text{ens}}(z|y_{i,t,k} + U_p y_{i,t})$.
  end for
end for
We end up with N new $f_{p,t}^{\text{ens}}(y_{i,t})$ for p-predictions (for each p of interest) for each i, t by which new p-portfolios are sorted.

Lemma 2. *Let $M_* = UDU'$ be the eigenvalue decomposition of M_* and U_p the matrix whose columns are the top p eigenvectors of M_*. Suppose that the data $x \sim N(\mu, \Sigma)$ is normally distributed. Let $f_p(y) : \mathbb{R}^p \to \mathbb{R}$ be defined via $f_p(y) = E[f(x)|U_p'x = y]$. Then,*

$$E[(f(x) - f_p(U_p'x))^2] \leq \lambda_1(\Sigma)\, \Lambda(M_*), \quad (17)$$

where

$$\Lambda_{-p}(M_*) = \sum_{i=p+1}^{d} \lambda_i(M_*). \quad (18)$$

Lemma 2 shows that $f_p(U_p'x)$ represents a good approximation to $f(x)$ as long as the residual variation $\Lambda_{-p}(M_*)$ is sufficiently small. The linear feature sparsity hypothesis is then simply a claim about the quality of the approximation of Lemma 2. The smaller the residual variation $\Lambda_{-p}(M_*)$, the more accurate the approximation is. To validate our analysis, we verify that $\Lambda_{-p}(M_*)$ is reasonably small for each CNNi model and each p of interest.

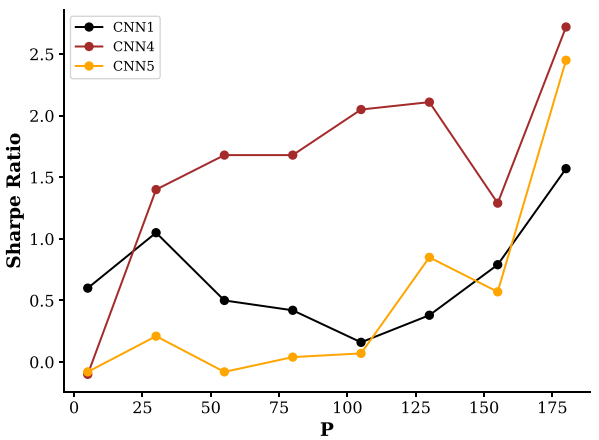


Figure 6 The figures above show the Sharpe Ratio of our H-L strategy (5) as a function of P, the number of principal features, based on the function $f_P(x)$ constructed using Algorithm 1. The experiment is run separately for each of the CNN1, CNN4, and CNN5 models.

Algorithm 1 describes the procedure of building the low-rank counterparts of the ensemble function $f^{\text{ens}}(x) = \frac{1}{K}\sum_{k=1}^{K} f(x; \hat{w}_\infty(w_o(k)))$. Using p-predictions obtained with $f_p^{\text{ens}}(x)$ counterparts, we build long-short portfolios and compare them by the Sharpe ratio.

Figure 6 shows the out-of-sample Sharpe ratios of this strategy as a function of p. We find a striking *virtue of feature complexity:* The out-of-sample performance is monotone increasing for $p > 50$, and the increase is slow. Even for $p = 100$, the Sharpe ratio is still significantly below the full model (corresponding to $p = 180$ linear features). Even for the low complexity CNN1 model, $p = 100$ recovers only about half of the full model Sharpe ratio, and the effect is even stronger for more complex models. This means that we need more than 100 linear features of the IV surface to capture the predictive relationships identified by the CNN models.

7 Conclusion

The remarkable growth of the factor zoo Feng *et al.* (2020), Bryzgalova *et al.* (2023) over the last few years has been accompanied by the development of machine learning methods for asset pricing Gu *et al.* (2020). As Kelly *et al.* (2021) and Didisheim *et al.* (2023) explain, this is no coincidence: Factor zoo is a natural consequence of *complexity*: A highly non-linear predictive relationship between returns and characteristics. The most naive and direct way of exploiting this complexity is to build large, unstructured non-linear models such as simple, fully connected neural networks of Gu *et al.* (2020) or the random feature models of Kelly *et al.* (2021) and Didisheim *et al.* (2023). While this approach works well with unstructured stock characteristics, it is unsuited for structured data, such as the IV surface. To deal with such data, one needs to develop tools and ML algorithms that exploit

the data structure optimally. In this paper, we take a step in this direction and propose a Convolutional Neural Network (CNN) architecture designed specifically to extract features of the IV surface that respect locality, as economic theory requires. We show that CNNs can successfully identify highly complex non-linear relationships that cannot be learned with naive, fully-connected networks. Importantly, we find that consistent with the existing evidence for image data Lakshminarayanan *et al.* (2017), the loss landscape of the CNN is extremely non-convex and is characterized by a very large number of local minima. All those minima contain information about returns. Exploiting them requires using an ensemble of CNNs. Gaining insights into the incremental information offered by the model as it converges to different local minima for other return prediction problems (including even simpler ones, with the fully connected networks of Gu *et al.* (2020)) is an important direction for future research.

Conventional wisdom based on the numerous manually constructed option characteristics suggests that a few linear features of the IV surface (e.g., level, slope, skew, and convexity) should fully summarize its predictive content. To test this "linear feature sparsity hypothesis," we introduce a novel object in financial machine learning, the *gradient outer product*, whose eigenvectors, the *principal linear features*, are natural analogs of principal components for machine learning Radhakrishnan *et al.* (2022). We find no evidence for linear feature sparsity and show that a very large number (more than 100) of linear features are necessary to explain the predictive content of IV, manifesting a very high *feature complexity*. Investigating principal linear features for other ML models and datasets might bring interesting novel insights into the different notions of sparsity in return prediction.

Appendix A Data Preprocessing

We get option implied volatility surfaces from the Option Metrics IvyDB. In IvyDB, each option chain is already normalized across expiration dates and deltas[18] (11 and 34 different values, respectively). We remove all rows with 10-day-to-expiration implied volatility values due to the high number of missing values at the beginning of the time frame. For each option implied volatility surface, we end up with a 2D matrix of size 10×34.

We link the CRSP dataset together with the CRSP delisting dataset, both available in WRDS, to take into account the delisting return as explained in subsection 3.1. As the Option Metrics database

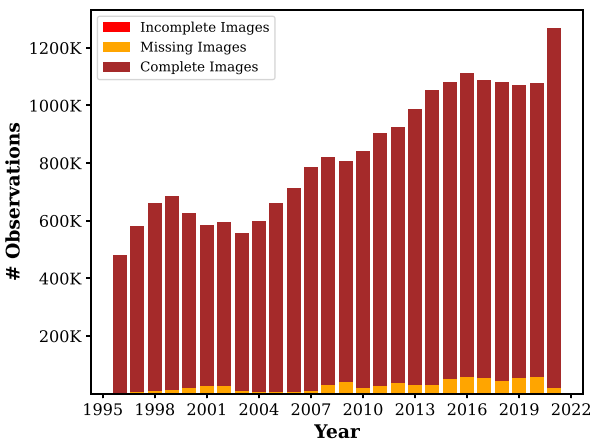


Figure 7 The figure above shows our dataset sample size. On the y-axis, we show the number of observations, while on the x-axis, we show the corresponding year. One observation is a single image obtained by stacking implied volatility surfaces for a given pair of stock and date. Incomplete Images are those observations (stock and date pair) where some implied volatility value is missing, given a particular expiration date and delta pair. Missing Images is the sample where the stock and date pair appear in the dataset, but have all empty values. Complete Images are the correctly built images where the implied volatility surfaces are complete.

uses its own security identifier and CRSP uses the PERMNO to identify an asset uniquely, we merge these datasets thanks to a linking dataset provided by WRDS. From the linking dataset, we remove "bad" entries:

(1) If an entry has a *score* lower than 1 is removed from our sample. A *score* of 1 means a 100% of mapping confidence.
(2) Each entry has a starting date (*sdate*), ending date (*edate*), Option Metrics ID (*secid*), and the PERMNO. We double-sort the dataset using *sdate* and PERMNO. If two consecutive entries have the same PERMNO, but the first (preceding) row has *edate* higher than the second (following) row's *sdate*, then the PERMNO is removed from our sample as well.

Appendix B Additional Results

B.1 Long-short portfolio performance

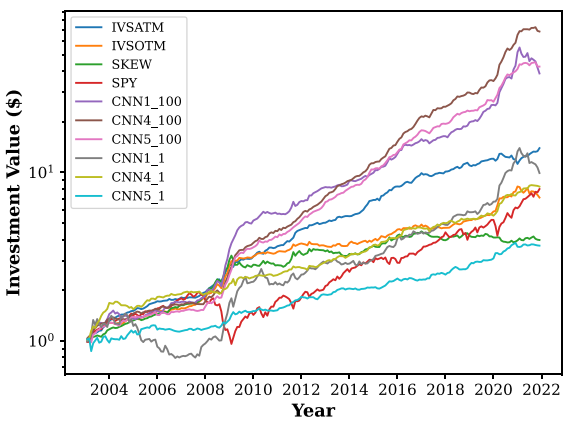


Figure 8 Cumulative Returns of the market neutral (long-short) strategy (5) for a full universe of stocks. CNN*i*_*K* refers to the ensemble of *K* CNN*i* models. IVSATM, IVSOTM, and SKEW are the top three option characteristics-based factors from Neuhierl *et al.* (2022) in terms of Sharpe ratio, and SPY is the S&P500 ETF.

B.2 Long-only portfolio performance

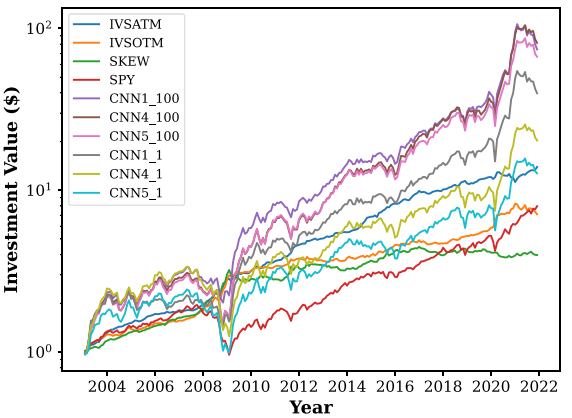


Figure 9 Long-only strategy (7). Cumulative Returns of (7) for the full universe of stocks. CNN*i*_*K* refers to the ensemble of *K* CNN*i* models. IVSATM, IVSOTM, and SKEW are the top three option characteristics-based factors from Neuhierl *et al.* (2022) and SPY is the S&P500 ETF.

Acknowledgment

We are grateful for the helpful comments from Fabio Trojani. Semyon Malamud gratefully acknowledges the financial support of the Swiss Finance Institute and the Swiss National Science Foundation, Grant 100018_192692. AQR Capital Management is a global investment management firm that may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR. This work was supported by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID sm81.

Endnotes

1 See Glebkin *et al.* (2023).
2 Formally, locality means that two close points on the IV surface have similar informational content.
3 The reason is that the dependence of CNN mean squared error on the CNN coefficients is highly non-convex.

As a result, the problem is many local minima, and it is impossible to predict to which local minimum the gradient descent will converge.

[4] Recent results of Jensen *et al.* (2022a) suggest that, by properly smoothing positions, it is possible to reduce turnover and preserve the bulk of the performance of machine learning models such as ours. We leave this important direction for future research.

[5] More generally, several recent papers have questioned the principle of sparsity in economic modeling, see, e.g., Giannone *et al.* (2021).

[6] See, also, Jiang *et al.* (2020) for similar results based on physical measures of skewness; and Kozhan *et al.* (2013), Schneider and Trojani (2019), and Orlowski *et al.* (2020) for studies of aggregate measures of skewness.

[7] This is likely because so-called "weeklies" (options with weekly expiration cycles) were only introduced in 2005 and gained popularity after the Global Financial Crisis.

[8] See Appendix A for a detailed description of CRSP and OM datasets, the dataset linking procedure, and some additional statistics.

[9] 1.34% of the total number of return observations.

[10] One may wonder whether information about the daily dynamics of implied volatility surfaces in the preceding month t contains predictive information about $R_{i,t+1}$ and whether we are neglecting this information by keeping only month-end $IV_{i,t}$. We have tried to incorporate additional information about daily IV dynamics in the preceding month in our predictive models, and these experiments suggested that lagged IV information during the month does not add predictive power.

[11] We also report that using a rolling window of five years, which allows us to account for potential non-stationarity, leads to a performance drop.

[12] See also Jiang *et al.* (2022).

[13] In particular, we use an increasing number of convolution filters in each block, i.e., 16, 32, 64, and 128.

[14] Weights are set in proportion to stock market capitalization, with market caps winsorized at 80% NYSE percentile as described in Jensen *et al.* (2022b).

[15] Restricting the analysis to linear transaction costs abstracts from important price impact considerations, see Jensen *et al.* (2022a).

[16] These estimates are very conservative. According to D'avolio (2002), most stocks have short-selling costs below 1% per annum.

[17] One common way of detecting sparsity in a non-linear model is by measuring feature importance. If most features get a low importance score, then one can conclude that the model is effectively sparse. The standard way of measuring feature importance is based on the Shapley value, see, e.g., Lundberg and Lee (2017). This approach is designed to measure the importance of a given feature that is expected to contain predictive information. This is the case for standard stock characteristics such as momentum, value, reversal, etc. (see, e.g., Jensen *et al.* (2022a)). However, in the case of the IV surface, the features are individual IVs. Hence, it is unlikely that a given IV for a specific (moneyness, maturity) combination contains distinct predictive information, and removing just one point on the IV surface will likely not impact model performance.

[18] In other words, each option implied volatility surface in IvyDB is interpolated across moneyness and maturities.

References

Amaya, D., Christoffersen, P., Jacobs, K., and Vasquez, A. (2015). Does Realized Skewness Predict the Cross-Section of Equity Returns?," *Journal of Financial Economics* **118**, 135–167.

An, B.-J., Ang, A., Bali, T. G., and Cakici, N. (2014). "The Joint Cross Section of Stocks and Options'," *The Journal of Finance* **69**, 2279–2337.

Andersen, T. G., Fusari, N., and Todorov, V. (2015). "The Risk Premia Embedded in Index Options'," *Journal of Financial Economics* **117**, 558–584.

Ang, A., Chen, J., and Xing, Y. (2006). "Downside Risk'," *Review of Financial Studies* **19**, 1191–1239.

Avramov, D., Cheng, S., and Metzker, L. (2021). "Machine Learning Versus Economic Restrictions: Evidence From Stock Return Predictability," *Available at SSRN 3450322*.

Bali, T. and Murray, S. (2019). In Search for a Factor Model for Optionable stocks.

Bali, T.G. and Hovakimian, A. (2009). "Volatility Spreads and Expected Stock Returns'," *Management Science* **55**, 1797–1812.

Bali, T. G., Chabi-Yo, F., and Murray, S. (2022). "A Factor Model for Stock Returns Based on Option Prices," *Available at SSRN 4071995*.

Baltussen, G., Bekkum, S., and Grient, B. (2018). "Unknown Unknowns: Uncertainty About Risk and Stock returns'," *Journal of Financial and Quantitative Analysis* **53**, 1615–1651.

Banz, R. W. and Miller, M. H. (1978), "Prices for State-Contingent Claims: Some Estimates and Applications'," *Journal of Business* 653–672.

Begin, J.-F., Dorion, C., and Gauthier, G. (2020). "Idiosyncratic Jump Risk Matters: Evidence From Equity Returns and Options'," *Review of Financial Studies* **33**, 155–211.

Belkin, M. (2021). "Fit Without Fear: Remarkable Mathematical Phenomena of Deep Learning Through the Prism of Interpolation," *Acta Numerica* **30**, 203–248.

Bollerslev, T. and Todorov, V. (2014). "Time-Varying Jump Tails', *Journal of Econometrics* **183**, 168–180.

Bollerslev, T., Todorov, V., and Xu, L. (2015). "Tail Risk Premia and Return Predictability'," *Journal of Financial Economics* **118**, 113–134.

Bollerslev, T., Tauchen, G., and Zhou, H. (2009), "Expected Stock Returns and Variance Risk Premia," *The Review of Financial Studies* **22**, 4463–4492.

Borovička, J., Hansen, L. P., and Scheinkman, J. A. (2016). "Misspecified Recovery'," *The Journal of Finance* **71**, 2493–2544.

Boyer, B., Mitton, T., and Vorking, K. (2010). "Expected Idiosyncratic Skewness'," *Review of Financial Studies* **23**, 169–202.

Brascamp, H. J. and Lieb, E. H. (1976). "On Extensions of the Brunn–Minkowski and Prékopa–Leindler Theorems, Including Inequalities for log Concave Functions, and With an Application to the Diffusion Equation," *Journal of Functional Analysis* **22**, 366–389.

Breeden, D. T. and Litzenberger, R. H. (1978a). "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business* 621–651.

Breeden, D. T. and Litzenberger, R. H. (1978b). "Prices of State-Contingent Claims Implicit in Option Prices'," *The Journal of Business* **51**, 621–651.

Bryzgalova, S., Huang, J., and Julliard, C. (2023). "Bayesian Solutions for the Factor Zoo: We Just Ran Two Quadrillion Models," *The Journal of Finance* **78**, 487–557.

Bryzgalova, S., Pelger, M., and Zhu, J. (2020). "Forest Through the Trees: Building Cross-Sections of Stock Returns," *Available at SSRN 3493458*.

Carhart, M. M., Carr, P., and Madan, D. B. (1997). "On Persistence in Mutual Fund Performance'," *The Journal of Finance* **52**, 57–82, Towards the theory of volatility trading, in R. Jarrow, ed., 'Risk book on.

Chen, L., Pelger, M., and Zhu, J. (2019). "Deep Learning in Asset Pricing," *arXiv preprint arXiv:1904.00745*.

Chen, Q., Roussanov, N., and Wang, X. (2021). "Semiparametric Conditional Factor Models: Estimation and Inference," *arXiv preprint arXiv:2112.07121*.

Chinco, A., Clark-Joseph, A. D., and Ye, M. (2019). "Sparse Signals in the Cross-Section of Returns," *The Journal of Finance* **74**, 449–492.

Chordia, T., Kurov, A., Muravyev, D., and Subrahmanyam, A. (2020). "Index Option Trading Activity and Market Returns'," *Management Science Forthcoming*.

Cochrane, J. H. (2011). "Presidential Address: Discount Rates," *The Journal of Finance* **66**, 1047–1108.

Conrad, J., Dittmar, R. F., and Ghysels, E. (2013). "Ex Ante Skewness and Expected Stock Returns'," *The Journal of Finance* **68**, 85–124.

Constantine, P. G., Dow, E., and Wang, Q. (2014). "Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces, *SIAM Journal on Scientific Computing* **36**, A1500–A1524.

Crego, J., Soerlie Kvaerner, J., and Stam, M. (2023). "Machine Learning and Expected Returns," *Available at SSRN 4345646*.

Cremers, M. and Weinbaum, D. (2010). "Deviations From Put-Call Parity and Stock Return Predictability," *The Journal of Financial and Quantitative Analysis* **45**, 335–367.

Davis, M. H. A. (2011). "The Dupire Formula," *Imperial College London, Finite Difference Methods Course Material*.

D'avolio, G. (2002). "The Market for Borrowing Stock," *Journal of Financial Economics* **66**, 271–306.

DeMiguel, V., Martin-Utrera, A. Nogales, F. J., and Uppal, R. (2020). "A Transaction-Cost Perspective on the Multitude of Firm Characteristics," *The Review of Financial Studies* **33**, 2180–2222.

Detzel, A., Novy-Marx, R., and Velikov, M. (2023). "Model Comparison with Transaction Costs," *The Journal of Finance*.

Dew-Becker, I. and Giglio, S. (2020). "Cross-Sectional Uncertainty and the Business Cycle: Evidence from 40 Years".

Didisheim, A., Ke, S., Kelly, B. T. and Malamud, S. (2023). "Complexity in Factor Pricing Models," *Swiss Finance Institute Research Paper*.

Duarte, J., Jones, C., and Wang, H. (2022). "Very Noisy Option Prices and Inference Regarding the Volatility Risk Premium".

Feng, G., Giglio, S., and Xiu, D. (2020). "Taming the Factor Zoo: A Test of New Factors," *The Journal of Finance* **75**, 1327–1370.

Feunou, B., Aliouchkin, R. L., Tedongap, R., and Xu, L. (2019). "Loss Uncertainty, Gain Uncertainty, and Expected Stock Returns".

Feunou, B., Jahan-Parvar, M. R., and Okou, C. (2018). "Downside Variance Risk Premium'," *Journal of Financial Econometrics* **16**, 341–383.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). "Deep Ensembles: A Loss Landscape Perspective," *arXiv preprint arXiv:1912.02757*.

Frazzini, A. and Pedersen, L. H. (2014). "Betting Against Beta," *Journal of Financial Economics* **111**, 1–25.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies* **33**, 2326–2377.

Giannone, D., Lenza, M., and Primiceri, G. E. (2021). "Economic Predictions with Big Data: The Illusion of Sparsity," *Econometrica* **89**, 2409–2437.

Giglio, S., Kelly, B., and Xiu, D. (2022). "Factor Models, Machine Learning, and Asset Pricing," *Annual Review of Financial Economics* **14**.

Giglio, S. and Xiu, D. (2021). "Asset Pricing with Omitted Factors," *Journal of Political Economy* **129**, 1947–1990.

Glebkin, S., Malamud, S., and Teguia, A. (2023). "Illiquidity and Higher Cumulants," *The Review of Financial Studies* **36**, 2131–2173.

Glorot, X. and Bengio, Y. (2010). "Understanding the Difficulty of Training Deep Feedforward Neural Networks, in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Gu, S., Kelly, B., and Xiu, D. (2020). "Empirical Asset Pricing Via Machine Learning," *The Review of Financial Studies* **33**, 2223–2273.

Guijarro-Ordonez, J., Pelger, M., and Zanotti, G. (2021). "Deep Learning Statistical Arbitrage," *arXiv preprint arXiv:2106.04028*.

Han, Y., Liu, F., and Tang, X. (2020). *The Information Content of the Implied Volatility Surface: Can Option Prices Predict jumps*?' (Working Paper).

Han, Y., He, A., Rapach, D., and Zhou, G. (2019). "Expected Stock Returns and Firm Characteristics: E-Lasso, Assessment, and Implications," *SSRN*.

Harvey, C. and Siddique, A. (2000). "Conditional Skewness in Asset Pricing Tests," *Journal of Finance* **55**, 1263–1295.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). "... and the Cross-Section of Expected Returns," *The Review of Financial Studies* **29**, 5–68.

Hou, K., Xue, C., and Zhang, L. (2020). "Replicating Anomalies," *The Review of Financial Studies* **33**, 2019–2133.

Ioffe, S. and Szegedy, C. (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd International Conference on Machine Learning* (*ICML 2015*) **1**, 448–456.

Jensen, C. S., Lando, D., and Pedersen, L. H. (2019). "Generalized Recovery," *Journal of Financial Economics* **133**, 154–174.

Jensen, T. I., Kelly, B. T., Malamud, S., and Pedersen, L. H. (2022a). "Machine Learning and the Implementable Efficient Frontier," *Available at SSRN 4187217*.

Jensen, T. I., Kelly, B. T., and Pedersen, L. H. (2022b). "Is There a Replication Crisis in Finance?," *Journal of Finance*.

Jiang, J., Kelly, B., and Xiu, D. (2022). "(Re-)Imag(in)ing Price Trends," Technical report.

Jiang, L., Wu, K., Zhou, G., and Zhu, Y. (2020), "Stock Return Asymmetry: Beyond Skewness, *Journal of Financial and Quantitative Analysis* **55**, 357–386.

Johnson, T. L., and So, E. C. (2012). "The Option to Stock Volume Ratio and Future Returns, *Journal of Financial Economics* **106**, 262–286.

Jones, C., Mo, H., and Wang, H. (2018). "Do Option Prices Forecast Aggregate Stock Returns?"

Kadan, O. and Tang, X. (2020). "A Bound on Expected Stock Returns," *Review of Financial Studies* **33**, 1565–1617.

Kelly, B., Malamud, S., and Zhou, K. (2021). "The Virtue of Complexity in Return Prediction," *Swiss Finance Institute Research Paper*.

Kelly, B., Pruitt, S., and Su, Y. (2020). "Characteristics are Covariances: A Unified Model of Risk and Return," *Journal of Financial Economics*.

Kilic, M. and Shaliastovich, I. (2019). "Good and Bad Variance Premia and Expected Returns, *Management Science* **65**, 2522–2544.

Kingma, D. P. and Ba, J. (2014). "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*.

Kozak, S., Nagel, S., and Santosh, S. (2018). "Interpreting Factor Models," *The Journal of Finance* **73**, 1183–1223.

Kozak, S., Nagel, S., and Santosh, S. (2020). "Shrinking the Cross-Section," *Journal of Financial Economics* **135**, 271–292.

Kozhan, R., Neuberger, A., and Schneider, P. (2013). "The Skew Risk Premium in the Equity Index Option Market'," *Review of Financial Studies* **26**, 2174–2203.

Lakshminarayanan, B., Pritzel, A. and Blundell, C. (2017). "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles," *Advances in Neural Information Processing Systems* **30**.

Leippold, M., Wang, Q., and Zhou, W. (2022). "Machine Learning in the Chinese Stock Market," *Journal of Financial Economics* **145**, 64–82.

Leoni, G. (2017). *A First Course in Sobolev Spaces* (American Mathematical Soc.).

Lettau, M. and Pelger, M. (2020). "Factors That Fit the Time Series and Cross-Section of Stock Returns," *The Review of Financial Studies* **33**, 2274–2325.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). "Visualizing the Loss Landscape of Neural Nets," *Advances in Neural Information Processing Systems* **31**.

Lin, H. and Todorov, V. (2019). "Aggregate Asymmetry in Idiosyncratic Jump Risk,".

Liu, Y., Zhou, G., and Zhu, Y. (2020). "Maximizing the Sharpe Ratio: A Genetic Programming Approach," *Available at SSRN 3726609*.

Lundberg, S. M. and Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems* **30**.

Martin, I. W. R. and Wagner, C. (2019). "What is the Expected Return on a Stock?, *The Journal of Finance* **74**, 1887–1929.

McLean, R. D. and Pontiff, J. (2016). "Does Academic Research Destroy Stock Return Predictability?," *The Journal of Finance* **71**, 5–32.

Moritz, B. and Zimmermann, T. (2016). "Tree-Based Conditional Portfolio Sorts: The Relation Between Past and Future Stock Returns," *Available at SSRN 2740751*.

Muravyev, D., Pearson, N. D., and Pollet, J. M. (2022). "Anomalies and Their Short-Sale Costs," *Available at SSRN 4266059*.

Neuhierl, A., Tang, X., Varneskov, R. T., and Zhou, G. (2022). "Option Characteristics as Cross-Sectional Predictors," *SSRN Electronic Journal*.

Orlowski, P., Schneider, P., and Trojani, F. (2020). "On the Nature of Jump Risk Premia," Unpublished Manuscript, Swiss Finance Institute.

Pederzoli, P. (2020). "Skewness Swaps on Individual Stocks".

Radhakrishnan, A., Beaglehole, D., Pandit, P., and Belkin, M. (2022). "Feature Learning in Neural Networks and Kernel Machines That Recursively Learn Features," *arXiv preprint arXiv:2212.13881*.

Ross, S. (2015). "The Recovery Theorem'," *The Journal of Finance* **70**, 615–648.

Schneider, P. and Trojani, F. (2019). "Divergence and the Price of Uncertainty'," *Journal of Financial Econometrics* **17**, 341–396.

Schneider, P., Wagner, C., Zechner, J., and Sharpe, W. F. (2020). "Low-Risk Anomalies?'," *The Journal of Finance* **75**, 2673–2718.

Shankar, V., Fang, A., Guo, W., Fridovich–Keil, S., Ragan–Kelley, J., Schmidt, L., and Recht, B. (2020). "Neural Kernels Without Tangents," in *International Conference on Machine Learning*, 8614–8623, PMLR.

Soudry, D. and Carmon, Y. (2016). "No Bad Local Minima: Data Independent Training Error Guarantees for Multilayer Neural Networks," *arXiv preprint arXiv:1605.08361*.

Stilger, P. S., Kostakis, A., and Poon, S.-H. (2017). "What Does Risk-Neutral Skewness Tell Us About Future Stock Returns?'," *Management Science* **63**, 1814–1834.

Tang, X. (2019). "Variance Asymmetry Managed Portfolios".

Xing, Y., Zhang, X., and Zhao, R. (2010). "What Does the Individual Option Volatility Smirk Tell Us About Future Equity Returns?'," *Journal of Financial and Quantitative Analysis* 641–662.

Yan, Y. (2011). "Jump Risk, Stock Returns and Slope of Implied Volatility Smile'," *Journal of Financial Economics* **99**, 216–223.

Yao, Z., Gholami, A., Keutzer, K., and Mahoney, M. W. (2020). "Pyhessian: Neural Networks Through the Lens of the Hessian, in *2020 IEEE International Conference on Big Data (Big data)*, 581–590, IEEE.

Keywords: Volatility surface; convolutional neural networks (CNN); machine learning; cross-section of stock returns; deep ensembles; virtue of complexity

Online Appendix Text Available