Journal of Economic Dynamics & Control 155 (2023) 104725

Contents lists available at ScienceDirect

Journal of Economic Dynamics and Control

journal homepage: www.elsevier.com/locate/jedc

Machine learning goes global: Cross-sectional return predictability in international stock markets

Nusret Cakici [a], Christian Fieberg [b,c,d], Daniel Metko [e], Adam Zaremba [f,g,h,*]

[a] Nusret Cakici, Gabelli School of Business, Fordham University, 45 Columbus Avenue, Room 510, New York, NY 10023, USA
[b] City University of Applied Sciences, Bremen, Germany
[c] University of Luxembourg, Luxembourg, Luxembourg
[d] Concordia University, Montreal, Canada
[e] University of Bremen, Bremen, Germany
[f] Montpellier Business School, 2300. avenue des Moulins, Montpellier, 34185 Cedex 4, France
[g] Department of Investment and Capital Markets, Institute of Finance, Poznan University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland
[h] Department of Finance and Tax, Faculty of Commerce, University of Cape Town, South Africa

ARTICLE INFO

JEL classification:
C52
G10
G12
G15

Keywords:
Machine learning
Return predictability
International stock markets
The cross-section of stock returns
Forecast combination
Asset pricing
Firm size

ABSTRACT

We examine return predictability with machine learning in 46 stock markets around the world. We calculate 148 firm characteristics and use them to feed a repertoire of different models. The algorithms extract predictability mainly from simple yet popular factor types—such as momentum, reversal, value, and size. All individual models generate substantial economic gains; however, combining them proves particularly effective. Despite the overall robustness, the machine learning performance depends heavily on firm size and availability of recent information. Furthermore, it varies internationally along two critical dimensions: the number of listed firms in the market and the average idiosyncratic risk limiting arbitrage.

1. Introduction

A recent merger between machine learning and asset pricing research has raised hopes for a solution to the factor zoo problem found within stock markets. So far, the evidence has been promising. Studies from the U.S. and China have demonstrated that machine learning models can effectively explain and predict the cross-section of stock returns.[1] Nevertheless, the evidence from individual

* Corresponding author.
E-mail addresses: adam.zaremba@ue.poznan.pl, a.zaremba@montpellier-bs.com (A. Zaremba).

[1] The evidence from the U.S. is by far the most extensive, concentrating on both simple signals from machine learning algorithms (Rapach et al., 2013; Feng et al., 2018, 2023; Freyberger et al., 2020; Gu et al., 2020; Han et al., 2023; Heaton et al., 2017; Azevedo & Hoegner, 2023; Rapach & Zhou, 2020; Avramov et al., 2023; Bali et al., 2021; Kim et al., 2021), as well as on high-dimensional asset-pricing models (Kelly et al., 2019; Chen et al., 2023; Kozak et al., 2020; Lettau & Pelger, 2020a, 2020b; Gu et al., 2021). For China, see, Leippold et al. (2021) and Hanaeuer and Kalsbach (2022).

https://doi.org/10.1016/j.jedc.2023.104725
Received 16 March 2023; Received in revised form 2 July 2023; Accepted 13 August 2023
Available online 18 August 2023

markets may not necessarily hold elsewhere.[2] Return predictability depends on the region or country characteristics that affect input data and model performance.[3] Furthermore, the complexity of machine learning techniques—as well as the diversity of their design—augments the risk of data dredging. To alleviate these concerns, extensive out-of-sample evidence across many countries and models is necessary.

Against this backdrop, this study combines machine learning with asset pricing research to comprehensively examine cross-sectional return predictability in 46 countries around the world. Using three decades of data from CRSP and Compustat, we calculate 148 anomaly variables and then use them to feed a repertoire of machine learning algorithms. The selection of models encompasses 11 representative algorithms: ordinary least squares regressions (OLS), partial least squares (PLS), the least absolute shrinkage and selection operator (LASSO), elastic net (ENET), support vector machine (SVM), gradient boosted regression trees (GBRT), random forests (RF), feed-forward neural networks with one, two, or three hidden layers (NN1, NN2, NN3), and forecast combination (COMB). Our global sample comprises over 74 thousand companies, nine million return observations, and more than one billion monthly stock characteristics. With this data at hand, we scrutinize the models' performance—their predictive abilities, characteristic importance, and economic gains from portfolio implementation. Finally, we explore the drivers of machine learning returns across countries. We want to establish what determines the differences in their profitability around the world.

Our findings contribute in five crucial ways. First, we reevaluate return predictability with machine learning models in 46 individual stock markets. Taken as a whole, the models deal relatively well in predicting future performance. In consequence, machine learning forecasts can be forged into profitable investment strategies. Though their performance varies markedly around the world, every algorithm generates sizeable abnormal returns in most countries. Even the simplest OLS, which allegedly suffers from overfitting, produces impressive returns on par with other models. Interestingly, all machine learning strategies perform visibly better in developed markets than in emerging ones. Finally, most markets display Sharpe ratios higher than the U.S., making a case for international diversification.

Our second contribution pertains to the superiority of the forecast combination model—the best performer in our study across numerous criteria. The COMB method averages predictions from all our individual models. The benefits of such an approach are well known in statistics (Bates and Granger, 1969; Clemen, 1989; Timmermann, 2006; O'Doherty et al., 2012), and the machine learning world is no different. While individual models have pros and cons, merging them reduces the forecast variance—leading to superior results. COMB is a clear winner of our model horserace. It produces not only the most accurate forecasts but also superior investment returns. When we pool all stocks in all countries together, a global long-short value-weighted forecast combination strategy earns 1.51% per month at an annualized Sharpe ratio of 1.49.

Third, our findings help to separate the wheat from the chaff in the global "factor zoo." Though hundreds of anomalies have been documented in stock returns, recent studies cast doubt on their validity (e.g., Harvey et al., 2016; Linnainmaa and Roberts, 2018; Hou et al., 2020). Thanks to the ability to digest multiple features at once, machine learning models help to pinpoint which stock characteristics really matter. Globally, the assorted machine learning models extract predictions from relatively similar variables. The top features correspond with well-known asset pricing phenomena—such as value, size, momentum, and reversal. Specifically, the most commonly selected characteristics are the current price ratio to the maximum price over the prior year, short-term reversal, performance mispricing factor, earnings to price, age, market equity, book-to-market equity, and three-, six-, and nine-month price momentum effects. Importantly, the precise contribution of different features varies considerably across countries—signaling that asset pricing does not follow uniform patterns everywhere. Put differently, the results from one market do not necessarily generalize to others.

Fourth, out results shed light on real-life setbacks challenging the real-life implementation of machine learning strategies. Concretely, portfolio performance depends on three essential factors: firm size, recent information, and high portfolio turnover. To begin with, machine learning works much better for small-cap stocks than for big-caps. Globally, machine learning strategies earn more than twice as high payoffs in small firms compared to big firms. The average six-factor model alphas in these two groups equal 2.37% and 0.99%, respectively. This regularity matches the arguments of Avramov et al. (2023), who state that machine learning strategies tend to extract abnormal returns from difficult-to-arbitrage stocks.

Next, the machine learning strategies hinge heavily on recent information. Typical machine learning tests assume the availability of closing prices from the date of portfolio formation. In reality, however, simultaneously fitting a model using a closing price and executing trades is hardly feasible. The transactions typically spread over some time following an investment decision. Unfortunately, machine learning models rely substantially on short-lived trading signals. Introducing a one-month skip period reduces the Sharpe ratio on the COMB strategy by 18.9% (on average across all markets and firm types). Among individual strategies, the dependence on recent information is particularly strong for complex algorithms, such as multilayer neural networks. Consequently, the additional skip period shrinks the average Sharpe ratio on the NN3 portfolios by 29.8%, turning it into the worst-performing strategy in our study. Finally—and in connection with the above—machine learning signals themselves tend to be short-lived. In consequence, the respective

[2] For example, Goyal and Wahal (2015) do not find robust evidence of the intermediate momentum effect of Novy-Marx (2012) outside the U.S. Jacobs and Müller (2020) show that the U.S. post-publication profitability decline (McLean & Pontiff, 2016) is non-existent in anomalies in other countries. Azevedo and Müller (2022) show that analysts' recommendations worldwide provide more value to the investors than earlier U.S. evidence has previously suggested.

[3] The example of studies of international heterogeneity in return predictability include, e.g., Barber et al. (2013), Chui et al. (2010), Gao et al. (2018), Jacobs (2016), Titman et al. (2013), Watanabe et al. (2013), Azevedo and Müller (2022), Cheon and Lee (2018), Docherty and Hurst (2018), Gao et al. (2018), Hollstein and Sejdiu (2020), and Cakici and Zaremba (2021b).

strategies require intensive trading and frequent portfolio reconstructions. The average monthly turnover typically falls between 80% and 140%, incurring potentially sizeable trading costs.

Last, our fifth contribution concerns the international heterogeneity in machine learning effectiveness. The performance of machine learning portfolios differs greatly across countries—the COMB Sharpe ratios can be as low as 0.20 in Austria and as high as 2.65 in Hong Kong. We identify two decisive factors driving this cross-country dispersion: the number of listed firms and average idiosyncratic risk. Fig. 1 illustrates the key punchline of our findings.

Fig. 1 demonstrates a compelling correlation: the performance of global machine learning strategies varies primarily based on two critical factors—the number of listed companies within a market (FIRMS) and the market's average idiosyncratic risk (IRISK).

Firstly, an increased number of listed companies contributes to the efficacy of machine learning models. Specifically, the COMB strategies in tercile of countries with the largest number of firms outperform those with the fewest stocks by 0.98% per month. The reason behind this lies in the breadth and depth of data available for analysis. With a more significant number of companies, machine learning models have access to a richer and more diverse dataset, facilitating improved training and tuning of these predictive algorithms. As machine learning models thrive on extensive data, a larger pool of companies naturally provides more opportunities for identifying patterns and exploiting inefficiencies. Furthermore, the number of firms shapes the factor structure in stock returns, as detailed by Bessembinder et al. (2021). Essentially, the diversity among firms in larger markets creates a more intricate network of return anomalies and potential trading opportunities. This diversity influences overall return predictability, enhancing the chance for machine learning models to detect anomalies and thereby increasing potential profitability.

Secondly, the machine learning strategies' performance is also strengthened by the level of idiosyncratic risk within the market. The difference in the mean returns on COMB strategies between top and bottom terciles of countries sorted by idiosyncratic risk equals 0.92%. Idiosyncratic risk, as established in asset pricing literature, is a notable proxy for limits to arbitrage, barriers that intensify mispricing and thereby amplify return predictability (Ali et al., 2003; Avarmov et al., 2021; Brav et al., 2010; McLean, 2010; Lam and Wei, 2011). High idiosyncratic risk often implies more significant mispricing due to higher limits to arbitrage. Stocks with more idiosyncratic volatility are more difficult to hedge, posing a higher risk for arbitrageurs. Moreover, it also relates to higher trading costs and uncertainty around company fundamentals, making mispricing exploitation and hedging even more challenging. As a result, the high idiosyncratic risk might also indicate lower informational efficiency, with stock prices possibly not fully incorporating all relevant data. This generates an environment ripe for machine learning models to exploit and capitalize on. Given their proficiency in handling vast amounts of complex data, they can leverage this informational gap, providing yet another avenue to boost returns.

Importantly, the role of the two variables indicated—*FIRMS* and *IRISK*—is remarkably robust. Their impact survive various tests—including cross-sectional regressions and country sorts and matters for all the considered machine learning models. They remain significant after controlling for each other, as well as in a multiple hypothesis testing framework.

Our findings relate to several strains of finance literature. First, we extend the discussion on machine learning applications to the cross-section of asset returns to international markets. Earlier evidence has concentrated mainly on the U.S., or several major markets, China or Western Europe.[4] Furthermore, a few papers scrutinized other asset classes—including government and corporate bonds (Bianchi et al., 2021; Bali et al., 2021), country and industry indices (Rapach et al., 2019; Cakici and Zaremba, 2022), commodities (Struck and Cheng, 2020; Rad et al., 2021), or currencies (Filippou et al., 2020).

To our knowledge, only four studies have—thus far—investigated the cross-sectional return predictability with machine learning in international markets: Tobek and Hronec (2021), as well as the working papers by Choi et al. (2022), Azevedo et al. (2022), and Hanauer and Kalsbach (2022). Tobek and Hronec (2021) concentrate on the benefits of using the U.S., global, and regional data to estimate the models. Their sample includes, on average, 2069 stocks per month from 23 developed markets (assigned into four regions). They find that whether data from other regions should be considered in model estimation depends on the region under consideration. Not only do we pursue a different research question, but we also follow a more holistic approach when it comes to data with 18,120 firms on average in the testing sample from both developed and emerging markets. In turn, the working paper by Choi et al. (2022) examines models based on 36 predictors from 31 countries (compared with 148 predictors in our case). Similar to Tobek and Hronec (2021), their focus is also on market integration, with tests of whether information extracted from the U.S. generate economic gains within international markets. The study reported in the working paper by Azevedo et al. (2022) is by far the largest—encompassing 44 countries. They examine the investment performance of machine learning strategies in a pooled global sample, with a particular emphasis on the post-publication decline. They do not explore other aspects of machine learning algorithms, such as prediction accuracy or variable importance, nor do they consider country-specific results. Finally, Hanauer and Kalsbach (2022) concentrate solely on emerging markets. Last but not least, none of the studies above investigate the sources of heterogeneity in international machine learning returns across markets.

Second, our study connects to the research on return predictability drivers across markets. Earlier research explored variables associated with limits to arbitrage, cultural traits, and market development in the context of individual predictors or anomalies (e.g., Barber et al., 2013; Chui et al., 2010; Gao et al., 2018; Jacobs, 2016; Titman et al., 2013; Watanabe et al., 2013; Azevedo et al., 2022; Cheon and Lee, 2018; Docherty and Hurst, 2018; Gao et al., 2018; Hollstein and Sejdiu, 2020; Cakici and Zaremba, 2022). Furthermore, our article corresponds closely with Bessembinder et al. (2021), who study the U.S. market to demonstrate that the number of

[4] The evidence from the U.S. market is rich and includes (but is not limited to), for example: Rapach et al. (2013), Feng et al. (2018, 2023), Kelly et al. (2019, 2023), Azevedo and Hoegner (2023), Chen et al. (2023), Freyberger et al. (2020), Gu et al. (2020, 2021), Han et al. (2023), Heaton et al. (2017), Kozak et al. (2020), Lettau and Pelger (2020a, 2020b), Rapach and Zhou (2020), Avramov et al. (2023), Bali et al. (2021), Kim et al. (2021), and Drobetz et al. (2021). For China, see Leippold et al. (2021) and Liu et al. (2022). For Europe: Drobetz and Otto (2021).

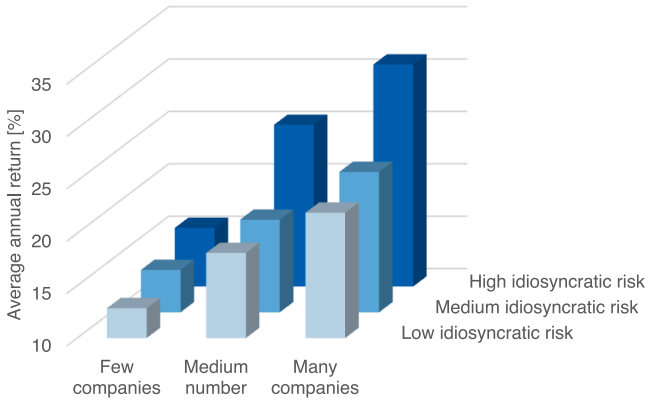

Fig. 1. Average Returns on Machine Learning Strategies in International Markets
The figure presents the average returns on forecast combination (COMB) machine learning strategies in markets grouped by the number of publicly listed firms (*FIRMS*) and average idiosyncratic risk (*IRISK*). The single-country long-short portfolios buy (sell) a value-weighted quintile of stocks with the highest (lowest) prediction from the COMB model. The strategies are applied in each of the 46 stock markets in our sample. The total study period is from January 1991 to December 2020; the testing period starts in January 2001. The markets are sequentially sorted into terciles: first based on *FIRMS* (*Few stocks, Medium stocks*, and *Many stocks*) and second based on *IRISK* (*Low idiosyncratic risk, Medium idiosyncratic risk*, and *High idiosyncratic risk*). We calculate the average monthly returns across all the strategies in each group and annualize them. The reported values are in percentages.

listed firms critically determines the factor structure. Against this background, we comprehensively dissect the determinants of the aggregate return predictability that is captured with machine learning models.

Third, in addition to dissecting the sources of heterogeneity in return predictability between countries, we also address differences in return predictability within countries. Specifically, we add to the evidence on the essential role of firm size in the magnitude of return predictability. Mounting academic literature documents that equity anomalies derive mainly from micro stocks (e.g., Hong et al., 2000; Fama and French, 2008, 2012; Novy-Marx, 2013; Cakici and Zaremba, 2021a, 2022); they also note that there is little predictability outside this firms segment (Hou et al., 2020; Hollstein, 2022). In the context of machine learning, Avramov et al. (2023) demonstrate that many algorithms derive their superior performance from difficult-to-arbitrage stocks. Müller and Schmickler (2021) show that the most powerful interactions, which typically propel machine learning profitability, originate in small and illiquid companies. In this context, we show that firm size plays an essential role in the profitability of machine learning strategies globally.

Last, our findings highlight the benefits of combining forecasts from individual models. In line with both theoretical and empirical evidence from statistics (Bates and Granger, 1969; Clemen, 1989; Timmermann, 2006; O'Doherty et al., 2012), merging individual predictions has been demonstrated to improve the accuracy of machine learning models (e.g., Rasekhschaffe and Jones, 2019; Bali et al., 2021; Azevedo et al., 2022). We provide convincing international evidence to support this view: a combination forecast works better than even complex machine learning models on a standalone basis.

The remainder of this paper proceeds as follows. Section 2 summarizes the data and methods. Section 3 presents the baseline empirical findings on prediction accuracy and characteristic importance. Section 4 focuses on the economic gains from portfolio implementations of the machine learning strategies. Section 5 concerns the international variations in machine learning returns. Finally, Section 6 concludes the study.

2. Data and methods

This section summarizes the data and methods. We begin by describing the data sources and sample preparation procedures. We then continue with a summary of stock characteristics that are used as model inputs. Last, we review the machine learning models employed in this study.

2.1. Playing field

Our sample covers 46 stock markets around the world. The study period is from January 1991 to December 2020; however, it starts later due to data unavailability for certain countries. Table 1 provides details of the sample structure and study periods across the different markets.

Return data for the U.S. market comes from CRSP. All return data for other countries, as well as all accounting data, is sourced from Compustat. Following a common approach in the international asset pricing literature (e.g., Fama and French, 2012, 2017), all market data is measured in U.S. dollars based on Compustat exchange rates. Consistent with this approach, the risk-free rate is proxied by the one-month U.S. Treasury bill rate.

Our sample is limited to common stocks, which are the primary securities of the underlying firms (as identified by Compustat). The companies are assigned to countries based on the country of their exchange. Moreover, each month, we exclude firms with a market

Table 1
Research sample.

Developed markets					Emerging markets				
Market	Start date	#total stocks	#average stocks	#features	Market	Start date	#total stocks	#average stocks	#features
Australia	Jan 1991	3518	974	133	Argentina	Jan 1996	147	60	103
Austria	Jan 1991	197	71	97	Brazil	Jan 1995	336	103	95
Belgium	Jan 1991	317	124	114	Chile	Jan 1994	270	119	121
Canada	Jan 1991	2979	889	147	China	Jan 1994	4271	1557	95
Denmark	Jan 1991	393	143	115	Colombia	Jan 1997	84	30	92
Finland	Jan 1991	269	108	110	India	Jan 1996	3906	1258	104
France	Jan 1991	1873	613	116	Indonesia	Jan 1996	828	313	106
Germany	Jan 1991	1752	589	103	Korea	Jan 1996	3337	1425	131
Hong Kong	Jan 1991	2799	1044	113	Kuwait	Jan 2005	235	136	77
Ireland	Jan 1991	121	41	107	Malaysia	Jan 1994	1399	765	121
Israel	Jan 1995	762	231	99	Mexico	Jan 1994	258	90	123
Italy	Jan 1991	782	256	108	Pakistan	Jan 1998	469	197	94
Japan	Jan 1991	5756	3319	144	Peru	Jan 1998	157	54	123
the Netherlands	Jan 1991	392	156	132	Philippines	Jan 1996	326	169	115
New Zealand	Jan 1991	302	98	99	Poland	Jan 1998	1060	313	105
Norway	Jan 1991	655	166	111	Qatar	Jan 2007	51	36	63
Portugal	Jan 1993	135	46	98	Russia	Jan 2005	630	153	108
Singapore	Jan 1991	1136	467	99	Saudi Arabia	Jan 2005	216	130	97
Spain	Jan 1991	430	143	113	South Africa	Jan 1994	991	290	126
Sweden	Jan 1991	1278	309	113	Taiwan	Jan 1996	2533	1216	112
Switzerland	Jan 1991	495	209	111	Thailand	Jan 1994	1081	467	107
UK	Jan 1991	5797	1685	137	Turkey	Jan 1996	534	277	103
USA	Jan 1991	18,817	5312	148	UAE	Jan 2005	136	65	67
Average	Apr 1991	2215	739	116	Average	Oct 1997	1011	401	104

The table presents the sample of countries that are covered in the study. *Start date* indicates the first monthly observation included in the calculations. *#total stocks* is the total number of unique stocks in a given market; *#average stocks* is the average monthly number of stocks in the sample; *#features* concerns the available stock characteristics.

capitalization below 5 million U.S. dollars. Last, the returns are winsorized each month at 0.1% and 99.9% to eliminate potential errors in the international data.

Once all the filters are applied, our sample comprises 74,210 companies, including 50,955 and 23,255 in developed and emerging markets, respectively. The total number of firm-month return observations is 8,928,138, and monthly stock characteristics exceed 1130 million.

2.2. Stock characteristics

With this dataset at hand, we closely reproduce 148 stock characteristics from Jensen et al. (2022). The selection encompasses the most prominent equity anomalies documented in finance literature. The detailed list is provided in Table A1 in the Internet Appendix. We calculate the variables following the Jensen et al. (2022) methodology—closely replicating all procedures.[5] For example, we compute all accounting variables using the most recent data—whether quarterly or annual—and assume that it becomes available four months after the end of a fiscal period.[6] Moreover, we replace any missing values with the cross-sectional median. Last, we transform each month's characteristics into ranks (based on country-specific rankings) and map them into an interval from −1 to 1.[7]

The precise sample size and span vary across countries (see Table 1 for details). While the study period in developed markets usually starts in 1991, the emerging markets coverage begins between 1994 and 2007. Furthermore, the average number of available firms and features differ. Whereas the developed markets cover 739 stocks—on average—with 116 characteristics, a typical emerging market comprises 401 stocks with 104 characteristics. The number of features is roughly similar to other large-scale tests in the current literature. These include McLean and Pontiff (2016) analyzing 97 anomalies, Green et al. (2017) exploring 94 signals, Gu et al., 2021) employing 94 characteristics, Jacobs and Müller (2020) exploring 241 predictors, Haddad et al. (2020) and Ehsani and Linnainmaa (2022) testing 50 factors, Jensen et al. (2022) replicating 153 variables, and Dong et al. (2022) using 100 anomalies.

[5] We are grateful to the authors for making their code available at https://github.com/bkelly-lab/ReplicationCrisis.

[6] One possible limitation of our international dataset may be differences in accounting standards across countries. We attempt to cope with this problem in two ways. First, we use the best possible data sources and calculation procedures identical to acknowledged international asset pricing studies (e.g., Jensen et al., 2022). Second, as we indicate in subsequent sections of the paper, we always estimate the models within each market separately. This guarantees that potentially inconsistent accounting data are not used in the same (pooled) estimation.

[7] For robustness, we experiment also with variable standardization instead of rank-mapping. This approach yields consistent returns, but the models' predictive efficiency and strategy performance are slightly weaker.

2.3. Machine learning models

We follow Gu et al. (2020) and employ a general additive prediction model to describe the relationship between the stock returns and characteristics:

$$r_{i,t+1} = E_t\left(r_{i,t+1}\right) + \varepsilon_{i,t+1}, \tag{1}$$

where $r_{i,t+1}$ is the excess return on security $i = 1, \ldots, N_T$ in month $t = 1, \ldots, T$. We calculate the expected return $E_t(r_{i,t+1})$ as a constant function of features available at t:

$$E_t\left(r_{i,t+1}\right) = g\left(z_{i,t}\right), \tag{2}$$

where $z_{i,t}$ denotes the vector of stock characteristics. It comprises up to 148 characteristics from Table A1. The function $g(z_{i,t})$ estimates the expected returns independently of any information from securities other than i or from periods before t. Notably, its exact form is left unspecified. Therefore, the flexible approximation functions depend on the family and can either be linear or nonlinear—as well as parametric or nonparametric.

All our models are designed to predict the true returns by minimizing the out-of-sample mean squared forecast error:

$$MSFE_{t+1} = \frac{1}{N_{t+1}} \sum_{i=1}^{N_{t+1}} \left(\widehat{\varepsilon}_{i,t+1}\right)^2, \tag{3}$$

where $\widehat{\varepsilon}_{i,t+1}$ is the individual prediction error for the stock i, and N_{t+1} is the number of stocks at $t+1$. In general, we seek the prediction model from a pool of candidates that displays superior forecasting accuracy.

We build on Gu et al. (2020), Leippold et al. (2022), and Bali et al. (2021) in order to select an array of representative machine learning models from finance literature. In consequence, we employ 11 different algorithms: ordinary least squares (OLS) regression, partial least squares (PLS), the least absolute shrinkage and selection operator (LASSO), elastic net (ENET), support vector machines (SVM), gradient boosted regression trees (GBRT), random forest (RF), and feed-forward neural networks with one to three layers (FFN1, FFN2, FFN3). Finally, motivated by the reasoning of Rapach et al. (2010) and Chen et al. (2023), we supplement this list with the forecast combination method (COMB); this averages the predictions from the 10 individual models. We provide more details on the models outlined above in Section B of the Internet Appendix.

We adopt the typical methods from machine learning literature to estimate the models, choose the hyperparameters, and evaluate the prediction performance. We calculate the models for each market separately. For each country, we split the total study period into three subperiods while maintaining the temporal order: the training period, comprising the first seven years of each sample (fixed window); the validation period, encompassing the next three years; and the testing period, the subsequent year. Overall, the testing sample encompasses the period from January 2001 to December 2020. To begin with, we use the training period to estimate the model's parameters subject to pre-specified model-specific hyperparameters. Next, we use the validation sample to tune the hyperparameters of the model. The aim of this optimization is to minimize the objective loss function. Last, we test the model's predictions using the subsequent 12 months following the validation period. Notably, the testing months are never included within the training or validation periods. As seen in the works of Gu et al. (2020) and Leippold et al. (2022), we re-estimate the models annually. The training, validation, and testing samples are rolled forward 12 months at each re-estimation.

3. Baseline empirical findings

This section reviews the primary evidence on the application of machine learning models within international markets. We first review the prediction performance of different algorithms. We then follow up by exploring the importance of specific stock characteristics.

3.1. Prediction performance of machine learning models

We start by comparing the predictive performance of different machine learning models in international markets. To provide a comprehensive picture, we calculate several different metrics.

3.1.1. Performance measures

We start with the classical predictive out-of-sample R^2 coefficient based on the pooled sample (R^2_{POS}), calculated as in Gu et al. (2020):

$$R^2_{POS} = 1 - \frac{\sum_{(i,t)\in T_3}\left(r_{i,t+1} - \widehat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in T_3} r^2_{i,t+1}}, \tag{4}$$

where $\widehat{r}_{i,t+1}$ and $r_{i,t+1}$ denote predicted and realized returns for stock i in month t, and T_3 indicates that we use data from the testing sample at re-estimation dates that never enter the training or validation samples. The measure is calculated based on the total sample of all return observations pooled across time and firms.

While R^2_{POS} is widespread in the finance literature, it has certain limitations. For example, it weights all observations equally, regardless of their importance for the investment portfolio. Assume that the number of listed firms in a market in two subsequent months grows from 100 to 200. Although each month's return would equally impact the portfolio value, the larger number of firms available in the second month would determine its stronger impact on the R^2_{POS} measure. Moreover, R^2_{POS} does not compare a model's predictions with any naïve forecast. To overcome these shortcomings, we also compute the cross-sectional out-of-sample R^2 coefficient (R^2_{CSOS}) by Han et al. (2023). This measure equally weights all periods in the sample and is computed in two steps.[8] Firstly, we calculate cross-sectional statistics $R^2_{CS,t}$ for each month:

$$R^2_{CSOS,t} = 1 - \frac{\sum_{i=1,..,T}^{N_t}\left[\left(r_{i,t+1} - \bar{r}_{i,t+1}\right) - \left(\hat{r}_{i,t+1} - \bar{\hat{r}}_{i,t+1}\right)\right]^2}{\sum_{i=1,..,T}^{n_t}\left(r_{i,t+1} - \bar{r}_{i,t+1}\right)^2}, \tag{5}$$

where $\bar{\hat{r}}_{i,t+1}$ and $\bar{r}_{i,t+1}$ are monthly average predicted and realized returns, respectively, based on all stock available at month t (N_t). Next, we calculate time-series averages of (5) to capture the predictability over the entire test period (T).

$$R^2_{CSOS} = \frac{1}{T}\sum_{t=1}^{T} R^2_{CS,t}, \tag{6}$$

The measure of Han et al. (2023) expresses the average proportional reduction in the monthly cross-sectional forecast errors relative to a benchmark naïve forecast that ignores information from stock characteristics.

Finally, in certain circumstances, the predictive R^2 coefficients may prove irrelevant for practical exercises such as portfolio sorts. Investors are typically interested in how effectively given measures rank stocks in accordance with the ex-post realized payoffs. This allows them to separate future market winners from losers. Nevertheless, in the popular R^2_{POS} and R^2_{CSOS} measures the correlation between forecasted and realized returns may be drowned in their variances—blurring the overall picture of the cross-sectional relationship (Coqueret, 2022). Consequently, investors may realize measurable economic gains even if the predictive R^2 is visibly negative (Kelly et al., 2023).

To cope with these issues, we supplement the R^2 measures with the simple average correlation coefficients over time. Specifically, each month we calculate the Pearson product-moment and Spearman rank-based correlation coefficients between the predicted and realized returns. Next, we calculate their averages across all months in the sample ($\bar{\rho}_P$, $\bar{\rho}_S$).

3.1.2. Evaluation results

Fig. 2 illustrates the distributions of prediction performance measures across countries. The models' accuracy varies substantially around the world. Observe the R^2_{POS} and R^2_{CSOS} in Panels A and B. Most coefficients range between −1.5% and 1%, but the dispersion is not uniform across models and measurement methods. The dispersion is more substantial for certain algorithms, such as OLS or NN1, but relatively lower for others, such as COMB. The variability in predictive performance bears important practical implications: the conclusions from one market do not necessarily generalize to others.

Table 2 provides their average accuracy measures across global, developed, and emerging markets.[9] Panel A reports the popular R^2_{POS} measure. Its global average values are frequently negative, although the specific values differ markedly across models and market types. The lowest average R^2_{POS} is recorded for OLS. This resembles the findings of Gu et al. (2020), who argue that a lack of regularization translates into a good in-sample fit, but disappointing out-of-sample predictions. The best individual model is NN3. This also matches findings from earlier seminal papers (Gu et al., 2020; Leippold et al., 2022), arguing that the superiority of neural networks comes from their ability to capture nonlinearities and interactions. The best method overall—in turn—is COMB. The combination method benefits from reducing the forecast variance associated with individual models. Hence, the reduced impact of uncorrelated prediction errors improves prediction accuracy. The observed superior performance of the COMB method echoes the earlier findings of Bali et al. (2021) from the U.S. market.

Interestingly, the R^2_{CSOS} measures in Table 2, Panel B, lead to partially similar conclusions than R^2_{POS}. COMB remains the best-performing technique. However, the precise pattern across other models is not always consistent. For example, this time, regularized regressions and tree models outperform the neural networks (NN1, NN2, NN3).

To complement the above findings with a more formal model comparison, Table A6 in the online appendix presents the tests of Diebold and Mariano (1995). While the pairwise differences in predictive accuracy often fall below commonly accepted statistical significance thresholds, the results confirm key patterns from the earlier analyses. The OLS model performs poorly and is often outperformed by more advanced techniques. On the other hand, the COMB approach proves to be particularly effective, beating a number of other methods. The differences are particularly evident for the global sample, which covers all 46 markets.

A closer look at the R^2 values in Table 2, Panels A and B, reveals another prominent pattern: the prediction performance is typically better in developed markets than in emerging ones. The strong predictability in developed markets may, at first sight, contradict the common narrative on inefficient emerging markets—if predictability originates from market inefficiencies. Nevertheless, this matches

[8] The measure is also discussed in Zaffaroni and Zhou (2022).

[9] For brevity, we limit the presentation in the main manuscript to the overall summary statistics. The detailed values of all measures for each country and method can be found in Tables A2 to A5 in the Internet Appendix.

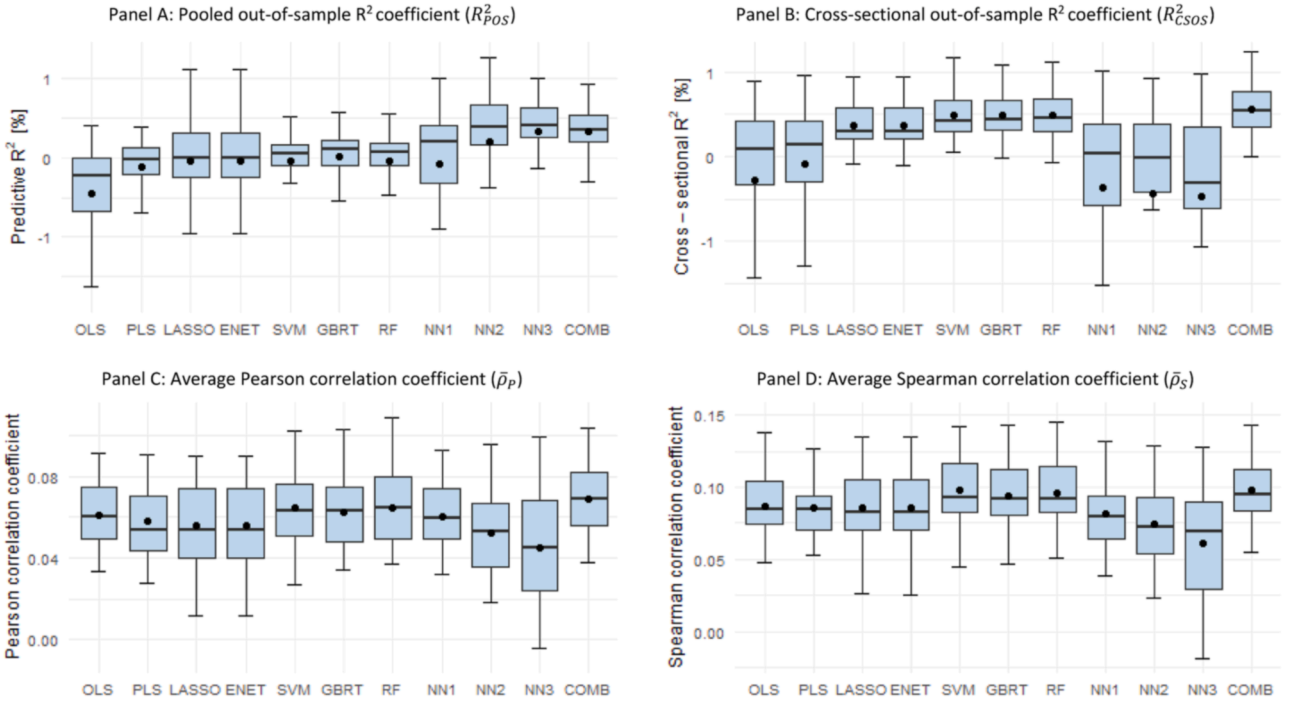

Fig. 2. Prediction Performance of Machine Learning Models

The figure exhibits box plots for the distributions of the prediction performance measures for different machine learning models (see Section 2.3) across the 46 markets covered in the study. Panel A presents the pooled out-of-sample R^2 coefficients (R^2_{OOS}) calculated as in Gu et al. (2020), and Panel B shows cross-sectional out-of-sample R^2 coefficients (R^2_{CSOS}) of Han et al. (2023). Panel C and D display the average monthly Pearson and Spearman correlation coefficients, respectively. The values in Panels A and B are in percentages. The body of each box represents the interquartile range, with its bottom and top marking the 25th and 75th percentile of the return distribution. The horizontal line in the middle of a box is the median. The whiskers indicate the maximum and minimum values. The black dot specifies the position of a mean. Outliers are excluded for the illustration quality. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

Table 2
Predictive R^2 coefficients for different machine learning models.

	OLS	PLS	LASSO	ENET	SVM	GBRT	RF	NN1	NN2	NN3	COMB
Panel A: Pooled out-of-sample R^2 coefficient (R^2_{POS})											
Global markets	−0.449	−0.117	−0.049	−0.050	−0.049	0.013	−0.041	−0.081	0.204	0.330	0.340
	(−3.72)	*(−2.09)*	*(−0.63)*	*(−0.64)*	*(−0.92)*	*(0.17)*	*(−0.77)*	*(−0.64)*	*(1.48)*	*(2.51)*	*(7.14)*
Developed markets	−0.211	0.076	0.191	0.192	0.113	0.148	0.118	0.132	0.261	0.164	0.388
	(−2.25)	*(2.34)*	*(3.35)*	*(3.36)*	*(2.85)*	*(4.04)*	*(3.25)*	*(1.35)*	*(1.29)*	*(0.77)*	*(9.52)*
Emerging markets	−0.687	−0.310	−0.289	−0.292	−0.211	−0.123	−0.200	−0.293	0.147	0.496	0.291
	(−3.27)	*(−3.50)*	*(−1.99)*	*(−1.99)*	*(−2.04)*	*(−1.24)*	*(−1.86)*	*(−1.15)*	*(0.96)*	*(3.55)*	*(3.46)*
Panel B: Cross-sectional out-of-sample R^2 coefficient (R^2_{CSOS})											
Global markets	−0.275	−0.091	0.364	0.361	0.483	0.478	0.480	−0.369	−0.447	−0.478	0.559
	(−1.57)	*(−0.77)*	*(10.20)*	*(9.81)*	*(14.19)*	*(13.85)*	*(14.13)*	*(−1.88)*	*(−1.54)*	*(−1.73)*	*(13.73)*
Developed markets	0.020	0.309	0.492	0.493	0.613	0.615	0.619	0.057	−0.258	−0.266	0.722
	(0.15)	*(4.01)*	*(9.43)*	*(9.46)*	*(13.16)*	*(13.41)*	*(13.35)*	*(0.43)*	*(−0.48)*	*(−0.50)*	*(13.61)*
Emerging markets	−0.571	−0.491	0.236	0.228	0.353	0.342	0.341	−0.795	−0.636	−0.690	0.396
	(−1.93)	*(−2.49)*	*(6.86)*	*(6.21)*	*(10.57)*	*(9.98)*	*(11.16)*	*(−2.26)*	*(−3.35)*	*(−5.37)*	*(9.71)*
Panel C: Average Pearson correlation coefficient ($\overline{\rho}_P$)											
Global markets	0.061	0.058	0.056	0.056	0.064	0.063	0.065	0.061	0.053	0.045	0.069
	(26.68)	*(24.40)*	*(20.58)*	*(20.31)*	*(27.15)*	*(25.15)*	*(25.52)*	*(26.38)*	*(17.41)*	*(10.84)*	*(29.15)*
Developed markets	0.068	0.068	0.067	0.067	0.073	0.072	0.075	0.069	0.068	0.069	0.079
	(26.03)	*(22.77)*	*(19.26)*	*(19.28)*	*(24.73)*	*(20.95)*	*(20.77)*	*(27.71)*	*(23.60)*	*(22.81)*	*(27.14)*
Emerging markets	0.054	0.048	0.045	0.045	0.056	0.054	0.055	0.052	0.037	0.022	0.059
	(17.84)	*(20.79)*	*(14.27)*	*(13.95)*	*(19.93)*	*(20.80)*	*(24.57)*	*(18.79)*	*(16.29)*	*(6.68)*	*(23.38)*
Panel D: Average Spearman correlation coefficient ($\overline{\rho}_S$)											
Global markets	0.086	0.085	0.086	0.086	0.098	0.094	0.096	0.082	0.074	0.061	0.098
	(25.10)	*(25.57)*	*(22.33)*	*(22.26)*	*(29.09)*	*(27.80)*	*(29.60)*	*(23.55)*	*(18.13)*	*(10.06)*	*(29.18)*
Developed markets	0.095	0.096	0.099	0.099	0.107	0.105	0.107	0.092	0.094	0.096	0.110
	(21.17)	*(22.46)*	*(21.29)*	*(21.36)*	*(24.76)*	*(24.23)*	*(25.68)*	*(20.33)*	*(23.50)*	*(23.30)*	*(25.67)*
Emerging markets	0.078	0.074	0.073	0.073	0.090	0.084	0.085	0.072	0.054	0.027	0.087
	(17.97)	*(19.80)*	*(15.28)*	*(15.15)*	*(21.44)*	*(21.11)*	*(23.50)*	*(18.03)*	*(15.20)*	*(5.74)*	*(22.37)*

The table presents the average prediction performance measures for different machine learning models (see Section 2.3) across the 46 global markets, 23 developed markets, and 23 emerging markets. Panel A presents the pooled out-of-sample R^2 coefficients (R^2_{OOS}) calculated as in Gu et al. (2020), and Panel B shows cross-sectional out-of-sample R^2 coefficients (R^2_{CSOS}) of Han et al. (2023). Panel C and D display the average monthly Pearson ($\overline{\rho}_P$) and Spearman ($\overline{\rho}_S$) correlation coefficients, respectively. The values in Panels A and **B** are in percentages. The numbers in parentheses are bootstrap *t*-statistics for cross-country averages. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

the empirical evidence of Jacobs (2016)—which notes that mispricing is at least as prevalent in developed markets as in emerging markets. One explanation may be the richer dataset; a broader cross-section of stocks and more variables allow for better algorithm optimization. On the other hand, the developed markets may exhibit a stronger factor structure, as they are typically populated by more firms (Bessembinder et al., 2021).

Panels C and D of Fig. 2 and Table 2 concentrate on the average cross-sectional correlation coefficients. Admittedly, their readings uncover several similarities to the R^2 measures. For example, once again, we observe an impressive performance of the COMB model. Interestingly, the correlation coefficients are consistently positive across all the methods in both developed and emerging markets. Globally, $\overline{\rho}_P$ ranges from 0.045 to 0.068 and $\overline{\rho}_S$ from 0.061 to 0.098 —paving the way for profitable machine learning strategies. Admittedly, while these values might be regarded as relatively low, they still suffice to generate measurable economic gains in all cases. This happens because the observed predictability is enhanced by canceling out idiosyncracies in a portfolio setting. In consequence, even methods with the lowest R^2, such as OLS, can deal with separating future winners from losers relatively well.

3.2. Which stock characteristics matter?

Having assessed the overall predictive efficiency of different models, we now investigate the relative importance of individual stock features. We aim to identify the essential determinants of the cross-section of stock returns in international markets while simultaneously accounting for the total "predictor zoo" within the system. To estimate the contribution of specific variables, we employ the approach of Kelly et al. (2019). Specifically, we calculate a predictor's variable importance (VI) as the reduction in predictive R^2 that results from setting all its values to zero while keeping the remaining model estimates fixed.

We begin by showing the ranking of the average VI for the 11 machine learning models in international markets. Fig. 3 illustrates the hierarchy of covariates by assigning the color gradient to characteristics. We first calculate the VI scores for each country and then average them across the 46 markets that are covered in our study. The VI values for each model are scaled so that their sum equals one. Dark (light) blue indicates the highest (lowest) characteristic importance. The features are sorted on their average VI across the 11 models.

Different machine learning models largely agree on the ranking of the leading variables. The most important feature is the ratio of the current price to the maximum price over the last year (*prc_highprc_252d*), closely followed by the short-term reversal (*ret_1_0*). The

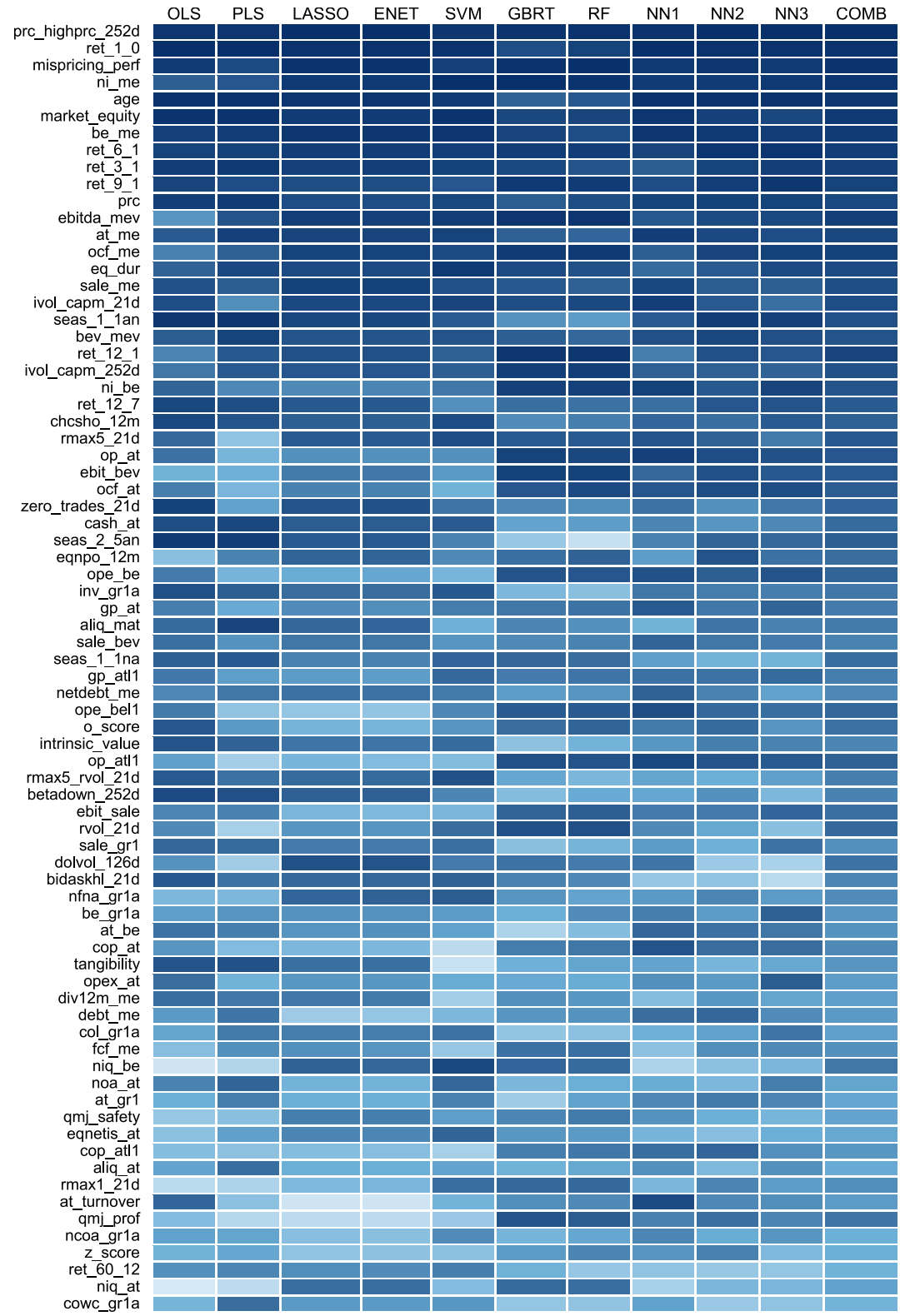

Fig. 3. Variable Importance

The figure presents the rankings of 148 return predictors that are considered in the study regarding their average total model contribution (see Section 2.3 for model description). The variable importance (VI) is calculated as the reduction of the overall OOS R^2 resulting from excluding a given variable from the model. First, we calculate the individual model contributions in each country separately and rescale them to sum to 1. Next, we

compute average contributions across the 46 countries in the sample. The color gradients indicate the rank of the variable importance; the dark blue (white) represents the most influential (least influential) predictors. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

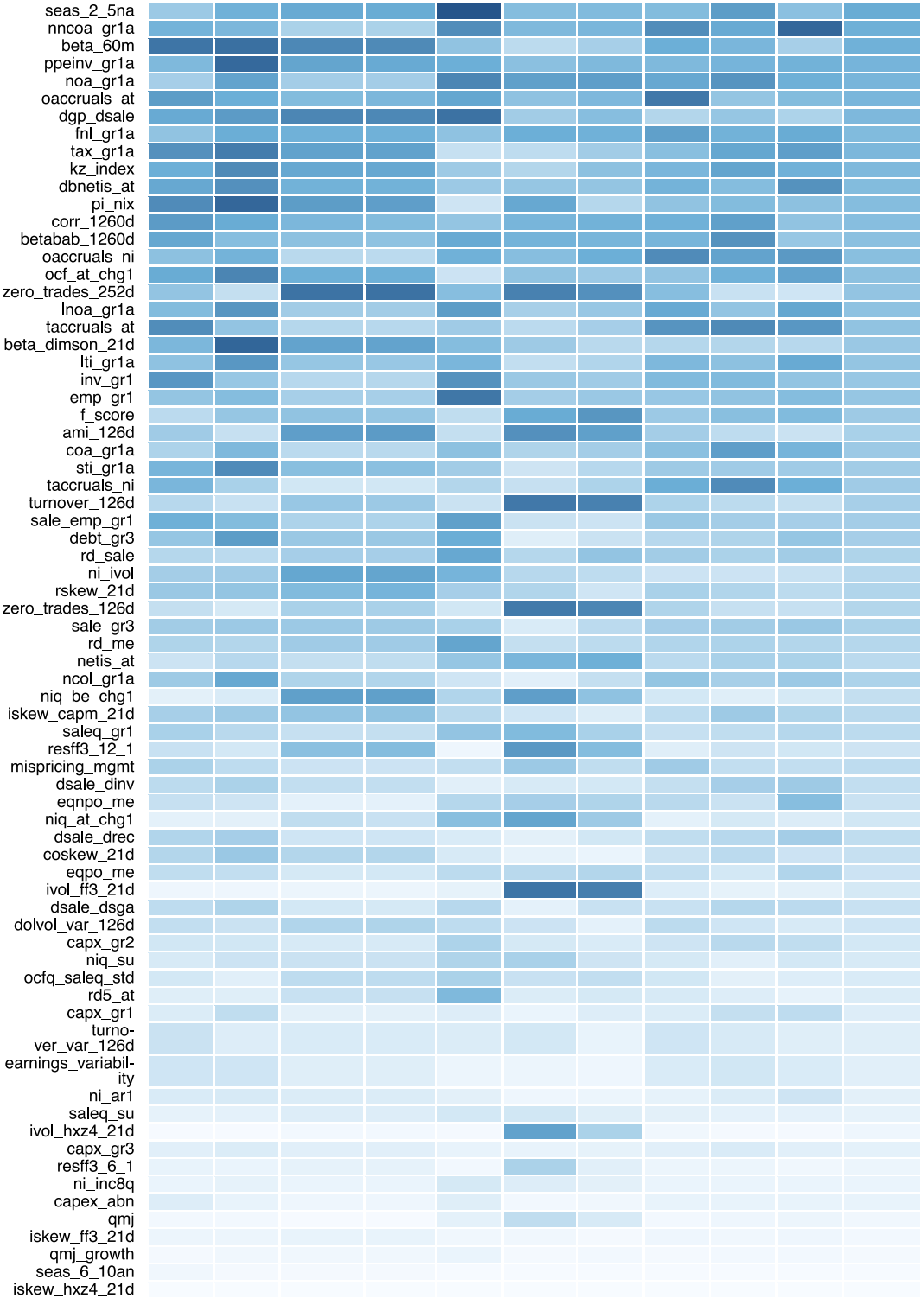

Fig. 3. (*continued*).

subsequent predictors encompass the performance mispricing factor (*mispricing_perf*), earnings to price (*ni_me*), age (*age*), market equity (*market_equity*), book-to-market equity (*be_me*), three-, six-, and nine-month price momentum (*ret_3_1, ret_6_1, ret_9_1*).

Interestingly, our feature selection partly matches the popular three- and four-factor models of Fama and French (1993) and Carhart (1997). However, it also specifies some other—even more important—variables, such as age or short-term reversal (which are not incorporated in these sparse asset pricing models). Our importance classification partly resembles the findings of Gu et al. (2020) for the U.S. market. For example, the indicators of past returns score high in both studies, and the short-term reversal—ranked the first variable in Gu et al. (2020)—is the second covariate in our classification. On the other hand, our results largely differ from Leippold et al. (2022)—who scrutinize the variable importance in Chinese equities. Unlike in China, we do not witness the essential role of market friction variables, such as illiquidity, number of zero trading days, or volatility of volume and turnover. Apparently, liquidity issues play a smaller role globally than in the emerging Chinese stock market.

Although the considered models similarly rank the variables, there are still some remarkable differences in how they assign relative importance. Figure A2 in the Internet Appendix shows the aggregate average importance of the top 5, 10, and 20 variables in different models. The penalized regression models (LASSO and ENET) are highly skewed toward the most essential features. Clearly, regularization leads to a focus on a relatively small number of characteristics. On the other hand, neural networks are visibly more democratic: they extract information from a broader set of variables.[10] Furthermore, the importance of top variables varies also to some extend across models and markets (see Figure A3 in the Internet Appendix). Once again, this observation highlights the risk of generalizations of findings in one market to a broader international context.

Finally, we are also interested in the relative importance of different categories of stock characteristics. Our earlier discussion focused on individual features. Nonetheless, certain groups of predictors—which share a similar economic intuition—may matter as a whole, although the contribution of individual variables is smaller. To elucidate this point, Fig. 4 depicts aggregate VI across different categories of variables. We follow the anomaly grouping of Jensen et al. (2022) in order to classify them into 13 broad baskets.

The crucial driver of stock prices around the world proves to be the value effect, closely followed by momentum anomalies. This observation corresponds with rich empirical evidence that picks value and momentum as the two most prominent and pervasive patterns in asset returns (Asness et al., 2013). The following categories include low-risk, quality, investment, and profitability factors. Most machine learning models rank the variable groups relatively consistently, though with some minor departures. For example, the tree methods (GBRT and RF) put a comparably larger weight on profitability and quality variables than the other models.

4. Machine learning portfolios

Our evidence has, thus far, confirmed the promising predictive abilities of machine learning methods. Earlier single-country studies (Gu et al., 2020; Leippold et al., 2022) argue that these forecasts can be forged into successful trading strategies. We now turn to portfolio analysis to see how these conclusions extend to various developed and emerging markets.

We begin with a general overview of the profitability of strategies based on different machine learning algorithms. Next, we take a closer look at the forecast combination strategy. Finally, we scrutinize the interplay between the firm size and machine learning returns.

4.1. Portfolio performance overview

We begin with country-level tests to overview the machine learning performance in the individual markets within the sample. Asset pricing studies usually sort stocks on a variable of interest into portfolios. Pursuing this reasoning, we group stocks into quintiles based on their monthly expected returns from the machine learning models. Next, we form long-short strategies that buy (sell) the top (bottom) quintile. In the baseline approach, we use value-weighting portfolios; however, we also check the equal-weighting scheme for robustness.[11] Finally, we examine the performance with the six-factor model of Fama and French (2018).[12] We apply these procedures in each of the 46 countries that are covered in this study.

Table 3 presents the average returns, Sharpe ratios, and alphas on the machine learning strategies across all countries in the sample—as well as across the developed and emerging markets only. Furthermore, Fig. 5 illustrates the performance distribution of different strategies across the 46 markets. The detailed results for individual countries are available in Tables A7 to A9 in the Internet Appendix.

A bird's eye overview of the outcomes reveals several interesting insights. To begin with, all machine learning models can be converted into profitable investment strategies. Even the techniques that exhibit negative R^2 values in Table 2, Panel A, produce impressive abnormal returns. This matches the arguments of Leitch and Tanner (1991), that the conventional error measures may not necessarily reflect the profit potential. The simple OLS, allegedly haunted by overfitting, performs surprisingly well. The average

[10] In practice, the sparsity of neural networks depends on their regularization. If regularized, they tend to derive information from a small group of variables.

[11] As seen in Jensen et al. (2022), we modify the pure value-weighting method and use a cap-value weighting approach to form balanced, yet tradeable, strategies. Specifically, we winsorize the market equity each month at the 80th percentiles to limit the impact of the largest companies.

[12] We closely follow Fama and French (2018) to calculate the model's underlying factors: market excess return, small minus big, high minus low, robust minus weak, conservative minus aggressive, and momentum. To assure apple-to-apple comparison, we derive the factor returns from our dataset that is described in Section 2.

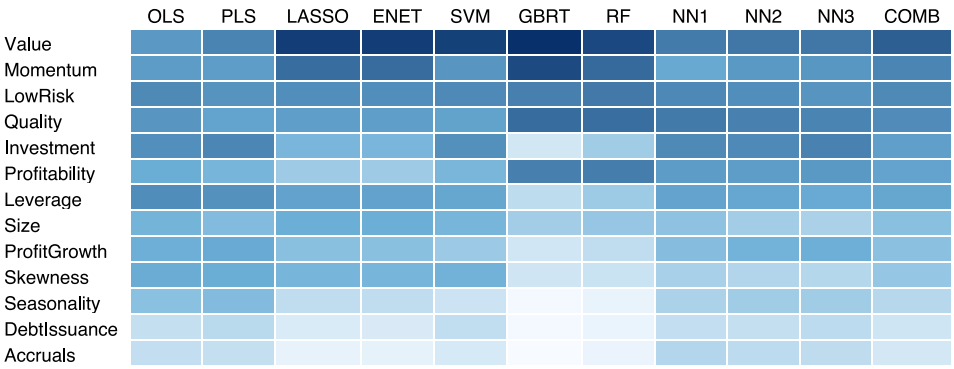

Fig. 4. Variable Importance per Category
The figure presents the rankings of 148 return predictors considered in the study regarding their average total model contribution aggregated within categories (see Section 2.3). The variable importance (VI) is calculated as the reduction of the overall OOS R^2 resulting from excluding a given variable from the model. First, we calculate the individual model contributions in each country separately and rescale them to sum to 1. Next, we aggregate them into clusters following the classification of Jensen et al. (2022) and using the methodology of Bali et al. (2021). Finally, we compute the average contributions across the 46 countries in the sample. The color gradients indicate the variable contribution; the dark blue (white) represents the most influential (least influential) categories of predictors. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

Table 3
Average performance of the machine learning strategies across countries.

	OLS	PLS	LASSO	ENET	SVM	GBRT	RF	NN1	NN2	NN3	COMB
Panel A: Mean monthly returns											
Global markets	1.20	1.25	1.24	1.24	1.41	1.33	1.43	1.22	1.11	0.93	1.52
	(11.14)	*(13.17)*	*(10.84)*	*(10.74)*	*(13.22)*	*(10.93)*	*(11.80)*	*(11.40)*	*(9.94)*	*(8.30)*	*(14.48)*
Developed markets	1.34	1.44	1.48	1.49	1.62	1.55	1.66	1.47	1.45	1.44	1.72
	(9.08)	*(11.50)*	*(10.21)*	*(10.14)*	*(10.54)*	*(10.48)*	*(10.58)*	*(11.57)*	*(12.77)*	*(12.88)*	*(11.05)*
Emerging markets	1.06	1.06	0.99	0.99	1.19	1.10	1.20	0.98	0.77	0.41	1.32
	(7.45)	*(8.83)*	*(6.47)*	*(6.42)*	*(9.55)*	*(6.33)*	*(7.30)*	*(6.60)*	*(5.17)*	*(3.86)*	*(10.55)*
Panel B: Sharpe ratios											
Global markets	0.99	0.92	0.83	0.84	0.99	0.86	0.89	1.03	0.95	0.73	1.09
	(10.40)	*(12.91)*	*(10.62)*	*(10.57)*	*(12.33)*	*(11.11)*	*(11.38)*	*(10.28)*	*(9.36)*	*(7.90)*	*(12.94)*
Developed markets	1.07	1.04	0.98	0.99	1.07	0.92	0.96	1.18	1.14	1.07	1.18
	(8.54)	*(11.28)*	*(9.69)*	*(9.74)*	*(9.90)*	*(9.73)*	*(9.55)*	*(9.80)*	*(9.54)*	*(9.88)*	*(9.49)*
Emerging markets	0.91	0.81	0.69	0.69	0.92	0.79	0.82	0.89	0.76	0.39	1.01
	(6.56)	*(8.08)*	*(6.19)*	*(6.13)*	*(7.90)*	*(6.60)*	*(7.11)*	*(5.95)*	*(5.31)*	*(3.68)*	*(8.88)*
Panel C: Six-factor model alphas											
Global markets	1.17	1.19	1.19	1.20	1.34	1.27	1.37	1.17	1.07	0.90	1.46
	(11.14)	*(12.54)*	*(10.40)*	*(10.31)*	*(12.63)*	*(10.89)*	*(11.17)*	*(10.96)*	*(9.10)*	*(8.45)*	*(13.78)*
Developed markets	1.31	1.37	1.43	1.44	1.54	1.49	1.60	1.41	1.41	1.39	1.66
	(9.26)	*(11.34)*	*(9.69)*	*(9.63)*	*(10.27)*	*(10.31)*	*(10.08)*	*(11.40)*	*(12.70)*	*(12.48)*	*(10.93)*
Emerging markets	1.04	1.02	0.95	0.95	1.14	1.06	1.15	0.94	0.72	0.42	1.25
	(7.41)	*(8.06)*	*(6.26)*	*(6.21)*	*(8.87)*	*(6.32)*	*(6.78)*	*(6.23)*	*(4.38)*	*(4.04)*	*(9.63)*

The table presents the performance of machine learning strategies (see Section 2.3) within international markets. For each country, each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The portfolios are value-weighted and rebalanced monthly. Panels A, B, and C concern the mean monthly returns, annualized Sharpe ratios, and alphas from the six-factor model of Fama and French (2018)—respectively. The mean returns and alphas are expressed in percentage terms. The reported values are average performance measures across the 23 developed markets, 23 emerging markets, and the pooled sample of 46 global markets. The numbers in parentheses are bootstrap *t*-statistics for cross-country averages. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

monthly alpha across all countries equals 1.20%; it is comparable with more sophisticated methods, such as dimension reduction techniques, regularized regressions, and neural networks. Furthermore, in terms of risk-adjusted performance, OLS can hardly be beaten by other machine learning strategies. With the average global Sharpe ratio of 0.99, it scores higher than all *individual* models—with the exception of NN1. The superior performance of OLS may suggest that the reputed benefits of the machine learning methods, such as accounting for interactions and nonlinearities, may be smaller than frequently believed. If the role of interdependencies and non-linear patterns in the data is subdued, the simplest prediction techniques—such as OLS—may also work very well.

Interestingly, our data do not reveal substantial benefits from attempts to battle the overfitting problem. While the dimension reduction techniques and penalized regressions translate into superior prediction efficiency (*vide* Table 2), they do not necessarily

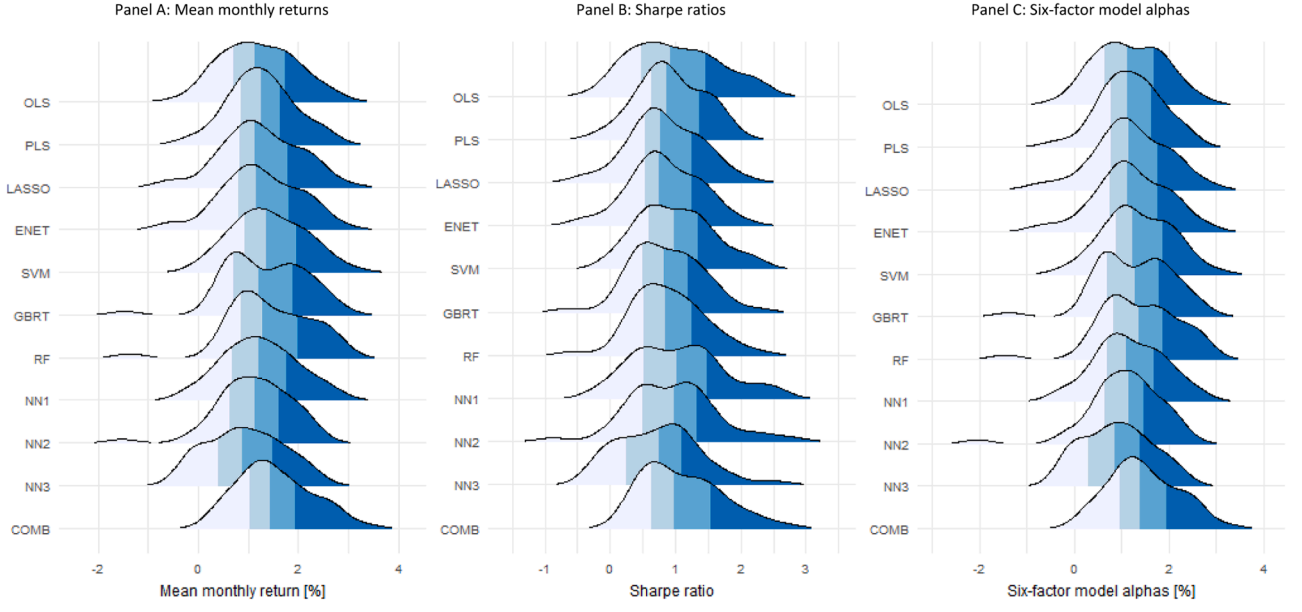

Fig. 5. Performance Distributions of Machine Learning Portfolios
The figure exhibits Gaussian kernel density plots for the mean monthly returns (Panel A), annualized Sharpe ratios (Panel B), and six-factor model alphas (Panel C) on different machine learning strategies (see Section 2.3) across the 46 stock markets covered in this study. We form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return for each country, each month. The portfolios are value-weighted and rebalanced monthly. The gradients represent different quartiles. The total study period is from January 1991 to December 2020; the testing period starts in January 2001. The mean returns and alphas are both expressed in percentage terms.

improve investment performance. For example, the global average alpha for the elastic net (ENET) equals 1.24%—closely resembling OLS. Moreover, its average global Sharpe ratio is inferior to the non-penalized regressions—amounting to 0.84 vs. 0.99 for OLS.

The support vector machines (SVM) and regression trees (GBRT, RF) seem to bring some progress, at least in terms of average returns and alphas. The best performing of these three models, RF, delivers the mean monthly return of 1.41%—and the associated average alpha is 1.34%. Nevertheless, the average annualized Sharpe ratios across all countries in the sample equals 0.99—still only on par with OLS.

The initially astonishing discrepancy between the Sharpe ratios and alphas may be intuitively understandable, given the mechanics of the tree models. This class is deemed to produce superior portfolio returns by seeking nonlinearities and interactions in stock returns. Meanwhile, the strongest interactions are typically associated with small and illiquid stocks (Müller and Schmickler, 2020)—which tend to be highly volatile. For example, the top features indicated in Fig. 3—such as short-term reversal, momentum, or valuation ratios—tend to be the strongest among small, illiquid, and volatile companies (e.g., Avramov et al., 2006; Jiang et al., 2005; Zhang, 2006; Fama and French, 2008, 2012). In line with this, Avramov et al. (2023) document that complex machine learning models extract profitability by difficult to arbitrage stocks—such as microcaps and distressed securities. Furthermore, the firm size premium (*market_equity*) also tends to be nonlinear and derives mainly from the smallest companies within the market (De Moor and Sercu, 2013). Thus, by pursuing these effects, the tree-based portfolio may implicitly allocate money to more volatile firms.

Finally, the results for feed-forward neural networks depend on the number of hidden layers. The shallowest version, NN1, has a Sharpe ratio of 1.03; thus, it produces a slightly better performance than OLS. The deeper versions of the networks, with two and three hidden layers, deliver gradually worse results. For example, the six-factor model alphas for NN1, NN2, and NN3 amount to 1.17%, 1.07%, and 0.90%—respectively. The superiority of "shallow" models over "deep" models is also noticed by Gu et al. (2020), who attributes it to the specificity of financial market data. Due to the low signal-to-noise ratio and the relative scarceness of data, the deep learning models prove less successful in asset pricing than—for instance—in bioinformatics or computer vision.

Thus far, the discussion has revolved around *individual* machine-learning methods. However, the real winner of the model horserace is the forecast combination (COMB). With the average global alphas of 1.46% and Sharpe ratios that amount to 1.09, COMB turns out to be the most profitable across all the methods. The superior performance of the COMB methods closely matches the findings of Gu et al. (2020) and Bali et al. (2021), who also found it overperforming in the U.S. market. COMB benefits from the superior prediction accuracy we have already demonstrated in Table 2. Averaging several models effectively reduces forecast variance, resulting in superior investment performance. While all models have their pros and cons, the forecast combination clearly surpasses them all.

Table 3 focused on performance differences across strategies. Nevertheless, even stronger variation is visible across countries. Fig. 5 illustrates the plots of the distributions of mean returns, Sharpe ratios, and alphas across all the markets in our sample. Clearly, the profitability is far from uniform. For example, the Sharpe ratios for the OLS strategy typically range from −0.5 to 2.5. Such a dispersion characterizes all strategies, as well as all performance measures.

Intriguingly, as seen in Table 3, an important dividing line for performance runs between the developed and emerging markets. Contrary to the common narrative, return predictability is visibly stronger in developed countries. This may seem surprising since inefficient emerging markets are frequently regarded as a reservoir of exploitable anomalies; however, the link between market maturity and mispricing is unequivocal (Jacobs, 2016). In developed markets, which are usually bigger, the machine learning algorithms may gain from richer datasets—which allow for better training. In line with this reasoning, the differences in profitability between developed and emerging markets are particularly pronounced for the most complex method. For example, NN3 in developed markets produces the alpha by 0.97 pp [=1.39%−0.42%] higher than in emerging ones. On the other hand, for the OLS, the analogous difference is only 0.27 pp [=1.31%−1.04%].

For robustness, we also reproduce the analyses above using equal-weighted portfolios. Table A10 and Fig. A1 in the Internet Appendix synthesize their performance, whereas Tables A11 to A13 report detailed results for each country. Overall, the findings are qualitatively consistent. The data evinces similar patterns: strong results of the simple OLS, superior profitability of the COMB strategy, and developed markets beating emerging ones. Nonetheless, one difference stands out: the equal-weighted performance visibly outperforms the value-weighted ones. The average global alphas increase by 0.28–0.58 pp, depending on the prediction model. The average Sharpe ratios are 35–50% higher. For example, the best-performing COMB strategy yields—on average (across the 46 markets)—a monthly alpha of 1.95% at a Sharpe ratio of 1.51.

This remarkable profitability of the equal-weighted portfolios relative to value-weighted ones is common to many anomaly-based strategies. The equal-weighting scheme allocates more capital to small stocks, where arbitrage is more challenging, and mispricing tends to be more pronounced. Many equity anomalies originate almost entirely from microcaps and can hardly be confirmed outside this segment (Hou et al., 2020). Likewise, the machine learning models may benefit from stronger return predictability in the small firm segment. Furthermore, higher return dispersion among the small companies may mechanically increase the spread of portfolio returns.

4.2. The forecast combination strategy

Thus far, the most successful machine learning technique turns out to be the forecast combination (COMB). Hence, we take a closer look at its performance around the world. Table 4 displays the returns on long-short portfolios that buy (sell) the quintile of stock with the highest (lowest) expected returns. The strategies are cap-value weighted, and the robustness checks for an equal-weighted portfolio are reported in Table A14 in the Internet Appendix.

The zero-investment strategy delivers remarkable performance, producing sizeable and significant profits in almost all countries.

For example, in the U.S., the average monthly return equals 1.10% and the six-factor model alpha is 1.06%. Notably, the COMB strategy works particularly well in big and liquid countries. For instance, the U.K. exhibits an alpha of 2.33% at a Sharpe ratio of 1.94. The alpha for Germany, the biggest European economy, is 2.59%. The strategy also works well in both Asia and Oceania, with the monthly alphas for Hong Kong and Australia amounting to 2.47% and 3.21%—respectively.

Notably, the strategies in most markets outperform their U.S. counterpart. For example, the Sharpe ratios in 27 countries (≈60%) exceed the value of 0.77, which we observed in the U.S. From a practical perspective, it emphasizes the measurable gains from investing internationally.

Despite this impressive performance, the COMB strategy returns show substantial variation across countries. Overall, the Sharpe ratios range from 0.20 (Austria) to 2.65 (Hong Kong). Several countries exhibit low and insignificant returns; this is especially true in small or emerging markets like Colombia, Qatar, and Saudi Arabia. The strategy is least effective in Saudi Arabia, where the six-factor

Table 4
Performance of the combination strategies in international markets.

	μ	t-stat$_\mu$	SD	SR	α	t-stat$_\alpha$	ML
Panel A: Developed markets							
Australia	3.31	*(7.04)*	5.30	2.17	3.21	*(7.85)*	21.11
Austria	0.42	*(0.73)*	7.25	0.20	0.45	*(0.73)*	54.07
Belgium	1.14	*(3.03)*	5.20	0.76	0.98	*(2.66)*	23.99
Canada	1.51	*(3.66)*	5.45	0.96	1.39	*(3.26)*	27.80
Denmark	1.90	*(4.80)*	5.54	1.19	1.86	*(4.11)*	19.23
Finland	1.14	*(3.13)*	5.32	0.74	1.06	*(2.65)*	14.38
France	1.76	*(4.75)*	4.91	1.24	1.77	*(5.61)*	22.38
Germany	2.71	*(5.76)*	6.00	1.57	2.59	*(5.92)*	23.66
Hong Kong	2.54	*(11.14)*	3.32	2.65	2.47	*(10.37)*	12.27
Ireland	1.40	*(1.71)*	11.33	0.43	1.42	*(1.84)*	42.84
Israel	1.43	*(4.49)*	4.30	1.15	1.22	*(3.47)*	16.09
Italy	2.05	*(4.83)*	6.59	1.08	2.07	*(4.71)*	25.68
Japan	1.00	*(5.48)*	2.70	1.29	0.90	*(4.75)*	8.52
the Netherlands	0.56	*(1.44)*	5.72	0.34	0.54	*(1.35)*	22.31
New Zealand	2.11	*(8.50)*	4.04	1.81	2.18	*(8.18)*	14.85
Norway	2.49	*(6.15)*	5.96	1.45	2.48	*(6.13)*	17.06
Portugal	1.53	*(3.05)*	7.07	0.75	1.40	*(2.93)*	29.01
Singapore	1.83	*(8.10)*	3.61	1.76	1.91	*(7.89)*	24.86
Spain	1.36	*(3.22)*	6.11	0.77	1.28	*(2.79)*	31.13
Sweden	2.76	*(5.31)*	6.79	1.40	2.59	*(5.27)*	21.89
Switzerland	0.98	*(2.66)*	5.38	0.63	1.04	*(2.76)*	29.17
UK	2.63	*(6.48)*	4.69	1.94	2.33	*(6.10)*	22.31
USA	1.10	*(2.72)*	4.95	0.77	1.06	*(3.16)*	21.29
Panel B: Emerging markets							
Argentina	1.31	*(2.83)*	6.96	0.65	1.42	*(2.82)*	18.97
Brazil	0.83	*(2.23)*	5.39	0.53	0.86	*(2.25)*	17.90
Chile	0.52	*(2.47)*	3.32	0.54	0.47	*(2.29)*	11.67
China	1.87	*(5.59)*	4.05	1.60	1.68	*(5.76)*	12.22
Colombia	0.47	*(1.37)*	5.00	0.33	0.28	*(0.76)*	23.50
India	1.90	*(4.81)*	5.40	1.22	1.95	*(5.34)*	29.35
Indonesia	1.49	*(3.35)*	5.26	0.98	1.40	*(3.51)*	18.96
Korea	2.66	*(7.18)*	4.04	2.28	2.59	*(6.66)*	9.85
Kuwait	1.26	*(3.49)*	3.27	1.34	1.44	*(3.40)*	7.74
Malaysia	2.14	*(9.53)*	3.58	2.07	2.08	*(8.13)*	12.58
Mexico	0.68	*(2.41)*	4.54	0.52	0.61	*(2.22)*	10.61
Pakistan	1.28	*(2.75)*	5.39	0.82	1.28	*(2.43)*	23.47
Peru	1.19	*(2.81)*	5.05	0.82	1.20	*(2.66)*	22.56
the Philippines	0.94	*(2.24)*	6.20	0.53	0.97	*(2.23)*	38.07
Poland	2.45	*(5.56)*	5.34	1.59	2.50	*(5.58)*	21.82
Qatar	0.60	*(0.65)*	4.90	0.42	0.34	*(0.40)*	13.36
Russia	1.29	*(2.08)*	4.55	0.98	0.99	*(1.62)*	11.40
Saudi Arabia	0.31	*(0.68)*	3.45	0.31	0.10	*(0.26)*	7.51
South Africa	1.74	*(6.23)*	3.83	1.57	1.60	*(5.98)*	14.40
Taiwan	1.46	*(5.41)*	3.77	1.34	1.42	*(5.05)*	13.36
Thailand	1.95	*(6.46)*	4.25	1.58	1.75	*(5.63)*	26.63
Turkey	0.98	*(2.38)*	6.03	0.56	0.91	*(2.24)*	22.11
UAE	1.04	*(2.01)*	5.82	0.62	0.97	*(1.63)*	24.66

The table presents the performance long-short strategies based on the forecast combination model (COMB). For each country, each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The portfolios are value-weighted and rebalanced monthly. The table reports the average monthly return (μ); the standard deviation of monthly returns (SD); annualized Sharpe ratio (SR); alpha from the Fama and French's (2018) six-factor model (α); and the maximum monthly loss, i.e., the most extreme negative return (ML).). μ, SD, α, and ML are expressed in percentage terms. The numbers in parentheses are Newey and West's (1987) adjusted t-statistics. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

model alpha equals 0.10%.

Overall, the results in Table 4 demonstrate a notable dispersion in the magnitude of machine learning profits. However, the nature of these differences can vary. On the one hand, from the standpoint of the Efficient Market Hypothesis (Fama, 1970), it might indicate lower informational efficiency in specific markets. For instance, higher market-wide limits to arbitrage may make it difficult to eliminate mispricing by sophisticate investors, leading to stronger return predictability. Additionally, this mispricing might be associated with local cultural idiosyncrasies or market development. On the other hand, the causes of abnormal returns could be purely mechanical, originating from factors such as a larger sample size that enables more effective training of the models. Section 5 delves further into the potential factors driving machine learning returns across different countries.

The last two rightmost columns of Table 4 report the maximum monthly loss—a statistic that matters from a practical perspective. The loss statistics appear manageable and normally fall between 10 and 30%. For example, the lowest values—not exceeding 10%—are recorded in Japan, Korea, and Kuwait. On the other hand, investors in Austria, Ireland, and the Philippines must have faced larger drawdowns, reaching even 30% to 50%.

Notably, many countries in our sample—especially among the emerging markets—impose significant restrictions on short selling. As a result, the long-short portfolios reported in Table 4 may not be entirely feasible. Therefore, Table A15 in the Internet Appendix reports returns separately for the top and bottom quintiles. Interestingly, unlike many individual anomalies (Stambaugh et al., 2012), the machine learning gains come mainly from long positions. In absolute terms, the abnormal returns of the top quintile considerably exceed those of the bottom quintile.

Nevertheless, while impressive, the performance of long-only strategies would not fully match that of their long-short counterparts. Globally, the average annualized Sharpe ratio for the top quintile is 0.72 versus 1.09 for the typical long-short strategy (see Table 3). The average long-only alpha is 1.29%, whereas the average long-short alpha is 1.46%, as shown in Table 3. In summary, while the long-only strategies that buy a quintile of the most promising stocks generate attractive risk-adjusted performance, including the short leg allows the risk-return profile to be further improved.

While the focus of Table 4 was on the returns in individual countries, it might also be worth investigating the performance of pooled international samples. In integrated markets, investors do not need to limit their scope to specific countries; however, they may apply bottom-up strategies directly in a broad global universe. To better understand the machine learning performance in this setting, Table 5 reports the results of the forecast combination (COMB) strategy in a pooled sample of global (46 countries), developed (23 countries), and emerging (23 countries) markets.

Table 5
Performance of the combination strategies in pooled global samples.

	Value-weighted portfolios					Equal-weighted portfolios				
	μ	SD	SR	α	Turn	μ	SD	SR	α	Turn
Panel A: Global markets										
Low	−0.16	5.98	−0.10	−0.18	67.00	−0.37	6.28	−0.20	−0.46	31.88
2	0.45	5.02	0.31	0.48	76.01	0.57	5.09	0.39	0.51	56.75
3	0.72	4.82	0.52	0.72	78.00	0.96	4.85	0.68	0.87	61.60
4	0.95	4.95	0.66	0.91	75.27	1.30	4.98	0.90	1.18	57.80
High	1.35	5.94	0.79	1.26	59.75	1.89	5.69	1.15	1.74	32.27
High-Low	1.51	3.50	1.49	1.44	126.75	2.26	3.10	2.52	2.19	64.15
	(5.64)			*(5.70)*		*(8.49)*			*(8.96)*	
Panel B: Developed markets										
Low	−0.21	6.26	−0.12	−0.21	64.20	−0.44	6.50	−0.23	−0.51	32.33
2	0.41	5.11	0.28	0.44	73.79	0.48	5.12	0.32	0.43	57.99
3	0.67	4.85	0.48	0.66	75.19	0.85	4.75	0.62	0.78	63.36
4	0.94	4.88	0.67	0.90	71.39	1.20	4.78	0.87	1.10	59.89
High	1.26	5.17	0.84	1.18	58.02	1.80	5.19	1.20	1.65	34.75
High-Low	1.46	2.75	1.85	1.39	122.22	2.23	2.78	2.79	2.16	67.08
	(6.27)			*(6.09)*		*(8.74)*			*(9.19)*	
Panel C: Emerging markets										
Low	0.03	6.73	0.01	−0.25	61.67	−0.15	6.56	−0.08	−0.50	31.37
2	0.67	6.27	0.37	0.46	72.29	0.83	6.06	0.47	0.53	55.15
3	0.86	6.12	0.49	0.65	74.15	1.10	5.94	0.64	0.78	59.61
4	1.01	6.53	0.54	0.76	71.75	1.26	6.17	0.71	0.86	55.93
High	1.49	7.44	0.69	1.21	58.23	1.77	6.94	0.89	1.37	31.35
High-Low	1.46	5.63	0.90	1.46	119.90	1.93	4.37	1.53	1.88	62.72
	(3.32)			*(2.73)*		*(5.81)*			*(4.52)*	

The table presents the performance quintile portfolios formed on the forecast combination model (COMB) implemented in pooled samples of stocks from 46 global markets (Panel A), 23 developed markets (Panel B), and 23 emerging markets (Panel C). Each month, we sort all stocks in the pooled samples into quintiles. *Low* (*High*) indicates the groups of firms with the lowest (highest) expected return according to the COMB model. The portfolios are equal- or value-weighted and rebalanced monthly. *High-Low* is the zero-investment long-short strategy that buys (sells) the *High* (*Low*) portfolios. The table reports the average monthly return (μ), the standard deviation of monthly returns (*SD*), annualized Sharpe ratio (*SR*), alpha from the Fama and French's (2018) six-factor model (α), and the portfolio turnover calculated as the average monthly change in holdings (Turn). μ, *SD*, α, and *Turn* are expressed in percentage terms. The numbers in parentheses are Newey and West's (1987) adjusted *t*-statistics. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

To form the reported portfolios, we first estimate the models in each country separately. Next, using the predictions from individual markets, we merge the country-specific samples into aggregate international samples. Last, we proceed with the standard portfolio sorts and performance evaluation—as seen in all earlier exercises.

Table 5, Panel A, demonstrates that the top value-weighted quintile of global stocks outperforms the bottom quintile by 1.51% monthly. The long-short strategy produces an alpha of 1.44% at a Sharpe ratio of 1.49. As seen in our earlier tests, the equal-weighted portfolios perform even better. The six-factor model alpha and Sharpe ratio on the spread portfolio are 2.19% and 2.52, respectively. This emphasizes the critical role of small firms in the magnitude of abnormal returns.

Table 5, Panels B and C, concerns the returns on the COMB strategy in aggregate, developed and emerging markets. In line with earlier observations in Tables 3 and 4, the outcomes confirm that the machine learning models in developed markets challenge their emerging market counterparts. The outperformance of developed market strategies manifests itself, particularly in risk-adjusted measures. For example, the Sharpe ratios on the value-weighted spread portfolio in developed and emerging markets equal 1.85 and 0.90—respectively. The equal-weighted portfolios generally earn-higher returns; however, the differences between developed and emerging countries' performance are even larger—with the respective Sharpe ratios in these market types equaling 2.79 and 1.53. The disparities between developed and emerging markets, as well between value- and equal-weighted portfolios, fit into our earlier interpretations concerning the critical role of sample size and firm capitalization in return predictability. The machine learning methods may benefit from more abundant data in developed markets that allow for more efficient model training. Furthermore, the focus on small stocks boosts the profitability of equal-weighted portfolios.

To complement the overview of international portfolios in Table 5, Fig. 6 plots their cumulative profits. The performance proves very stable, particularly for developed markets. There are no major drawdowns; furthermore, the swings following major market downturns—such as the global financial crisis in 2008—are limited. Additionally, the strategies exhibit no visible attenuation of return predictability due to investor learning or changes in market efficiency. This pattern complies with the findings of Jacobs and Müller (2020), who also observe no reliable decline in stock return predictability in recent years within international stock markets.

4.3. *Machine learning profitability and firm size*

Avramov et al. (2023) argue that machine learning models extract profitability from difficult-to-arbitrage stocks. Furthermore, Hou et al. (2020) and Hollstein (2022) demonstrate that most return predictability in financial markets is derived from the smallest firms. Whilst their role is accounted for, the profitability of characteristic-based portfolios visibly weakens. The firm size evinces itself as one of the key drivers of return predictability within equity markets.[13]

To scrutinize the role of firm value in generating returns by the machine learning strategies, we replicate our analyses in the subsample of small and big firms. We conduct these tests in three steps. First, we estimate the return forecasts for each stock in each country using our usual models and procedures. Second, each month, we split each market into halves by the median stock capitalization at t-1 and classify the firms above (below) the median as big (small). Third, we run our standard sorts into quintile portfolios in each of these size subsets.

Table 6 reports the mean returns and alphas on the value-weighted strategies that are implemented in big and small firms.[14] For brevity, our baseline analysis pertains only to the forecast combination (COMB) strategy—which aggregates all the individual models. In line with our earlier intuitions, there is a striking disparity in the long-short strategy performance in small and big companies. The mean raw and abnormal returns in the small-firm segment are considerably higher than in big firms. For example, small firms' monthly six-factor model alphas range from 0.51% (Peru) to 4.66% (Singapore)—with a global cross-sectional average of 2.37%. On the other hand, the worst and the best alphas in big firms are −0.79% (Qatar) and 2.09% (Australia)—with the cross-country average being 0.99%. Furthermore, the outperformance of small firms holds for developed and emerging markets alike; in both of these segments, the machine learning strategies normally earn two to three times more in small stocks than in large ones.

While Table 6 details the COMB strategy, Fig. 6 succinctly extends these analyses to all individual strategies. Specifically, it plots the distributions of performance measures (mean returns, Sharpe ratios, and alphas) for the machine learning strategies implemented in the big and small firms across the 46 markets considered. Their overview leads to consistent conclusions for the methods considered. While all the measures for all methods show certain dispersion across countries, the distributions for small stocks are shifted rightwards in each of the cases. In other words, the performance of the machine learning models is consistently stronger in small companies than in big ones; this applies to various methods and performance measures.

One additional insight from Fig. 7 is that the small firm strategies generally perform better but exhibit more sizeable cross-country variation in results. All the distributions are flatter and broader, signaling larger uncertainty around the possible strategy performance. In practice, this implies that the results for small stocks in one market are generalizable to a lesser extent than for the big stocks.

The superior performance of the machine learning strategies in small firms may derive from two different sources. On the one hand, it may originate from enhanced return predictability in small cap segments; higher limits to arbitrage, slowly traveling news and capital, and lower market efficiency may result in sturdier return patterns. Alternatively, high returns may be generated mechanically. Small firms are typically more volatile and exhibit higher cross-sectional dispersion of returns. Consequently, the return spreads

[13] For further studies covering the role of firm size in anomaly-baled return predictability—see, e.g., Hong et al. (2000), Fama and French (2008, 2012), Novy-Marx (2013), and Cakici and Zaremba (2021a, 2021b).

[14] An analysis of equal-weighted portfolios leads to qualitatively similar conclusions concerning the differences between small and big firms. Table A16 in the Internet Appendix contains the detailed results of these tests.

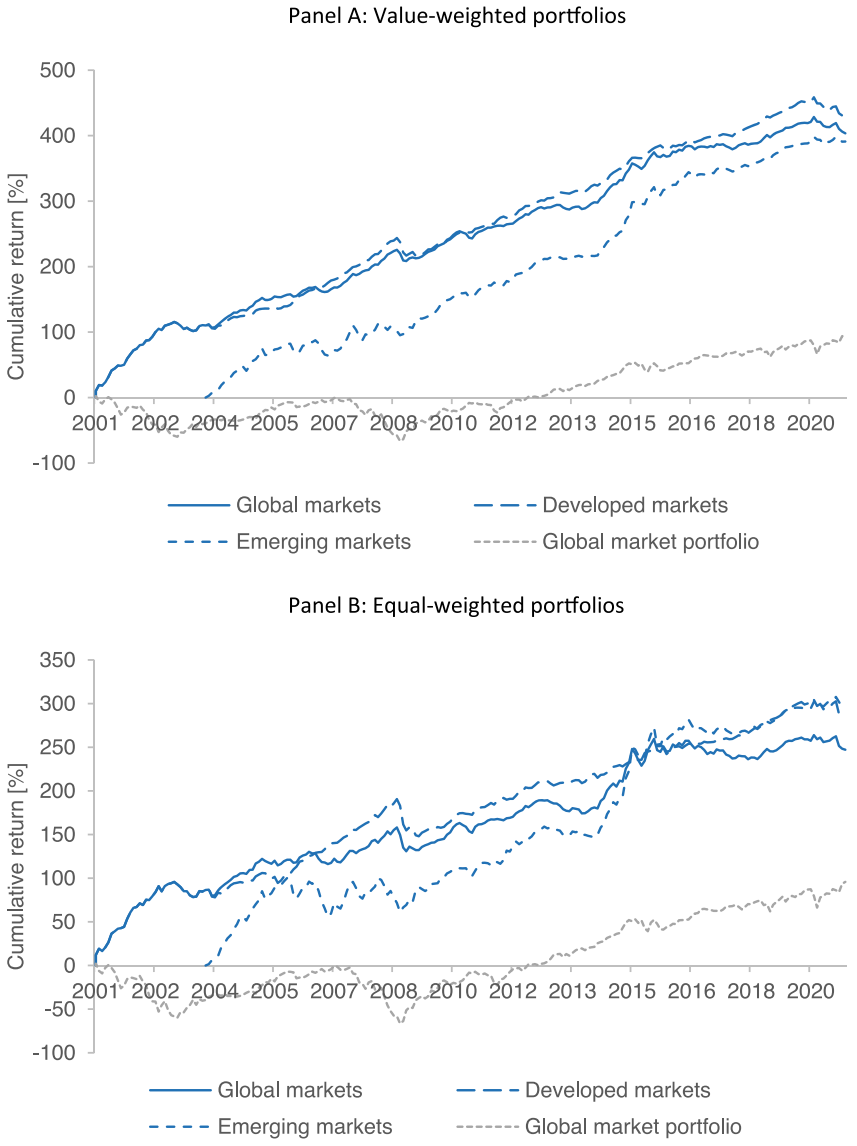

Fig. 6. Cumulative Returns on the Combination Strategies in Global Markets
The figure presents the cumulative returns through time on the zero-investment forecast combination strategies (COMB) implemented in pooled samples of stocks from 46 global markets, 23 developed markets, and 23 emerging markets. The displayed long-short strategies buy (sell) the quintile of stocks with the highest (lowest) expected return. The portfolios are value- or equal-weighted (Panels A and B, respectively) and rebalanced monthly. Additionally, the gray line represents the cumulative excess return on the global market portfolio, represented by the MSCI ACWI Index. The monthly returns are cumulated additively and reported in percentage terms. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

between extreme quantiles tend to be broader.

To shed light on the question above, we calculate the prediction accuracy measures from Table 2 for the subsamples of big and small stocks. We want to verify whether the superior returns on small firms are associated with stronger return predictability. For conciseness, we report these results in Table A17 in the Internet Appendix. The R^2 and correlation coefficients are typically higher for small stocks. For example, the average global R^2_{POS} equal 0.40 and 0.24 in small- and large-cap segments, respectively. The average global $\overline{\rho}_S$—which is perhaps the most informative measure in the context of portfolio sorts—equals 0.114 in the small-cap segment and 0.070 in the large-cap ones. Admittedly, we can observe certain variation across individual markets, but the overall pattern is clear: the stronger return predictability in small stocks boosts the profitability of machine learning strategies.

Table 6
Returns of the combination strategies in big and small firms.

	Small firms					Big firms				
	μ	t-stat$_\mu$	SR	α	t-stat$_\alpha$	μ	t-stat$_\mu$	SR	α	t-stat$_\alpha$
Panel A: Developed markets										
Australia	3.42	*(9.26)*	2.58	3.37	*(9.04)*	2.18	*(5.43)*	1.63	2.09	*(5.65)*
Austria	2.63	*(4.20)*	1.15	2.71	*(4.86)*	0.32	*(0.65)*	0.18	0.36	*(0.75)*
Belgium	2.12	*(4.39)*	1.04	2.17	*(4.64)*	0.51	*(1.40)*	0.36	0.43	*(1.08)*
Canada	3.37	*(8.83)*	2.31	3.12	*(8.05)*	1.04	*(2.79)*	0.70	0.97	*(2.40)*
Denmark	3.15	*(7.94)*	1.66	3.03	*(6.83)*	1.50	*(3.61)*	0.89	1.35	*(2.57)*
Finland	2.62	*(5.57)*	1.36	2.50	*(5.23)*	1.06	*(2.70)*	0.67	1.13	*(2.51)*
France	3.11	*(7.02)*	1.85	3.08	*(7.24)*	0.84	*(2.76)*	0.74	0.84	*(3.11)*
Germany	3.87	*(7.70)*	2.28	3.78	*(8.67)*	1.56	*(3.35)*	0.92	1.46	*(3.25)*
Hong Kong	3.46	*(8.41)*	2.34	3.28	*(8.03)*	2.07	*(8.59)*	2.02	2.01	*(8.31)*
Ireland	3.58	*(4.57)*	1.02	4.05	*(5.27)*	1.60	*(2.05)*	0.49	1.81	*(2.31)*
Israel	2.28	*(4.71)*	1.32	2.13	*(4.56)*	0.76	*(2.38)*	0.69	0.58	*(1.72)*
Italy	2.44	*(6.46)*	1.46	2.21	*(5.45)*	1.07	*(2.92)*	0.65	1.11	*(2.92)*
Japan	1.50	*(6.67)*	1.46	1.42	*(6.01)*	0.80	*(4.15)*	0.95	0.72	*(3.60)*
the Netherlands	1.79	*(2.83)*	0.74	1.79	*(2.77)*	0.38	*(1.02)*	0.23	0.43	*(1.06)*
New Zealand	3.02	*(6.55)*	1.50	3.06	*(6.95)*	1.31	*(4.95)*	1.07	1.25	*(4.47)*
Norway	3.24	*(5.42)*	1.39	3.30	*(5.74)*	1.75	*(4.84)*	1.09	1.64	*(4.41)*
Portugal	1.94	*(2.64)*	0.65	1.53	*(1.96)*	1.13	*(2.06)*	0.53	0.93	*(1.89)*
Singapore	4.35	*(11.11)*	2.97	4.66	*(11.86)*	1.29	*(5.43)*	1.19	1.27	*(5.15)*
Spain	2.18	*(4.28)*	1.08	2.19	*(4.15)*	1.07	*(3.05)*	0.72	1.04	*(2.82)*
Sweden	2.89	*(6.08)*	1.32	3.00	*(6.83)*	1.38	*(3.09)*	0.85	1.14	*(2.66)*
Switzerland	2.42	*(4.93)*	1.43	2.58	*(5.82)*	0.83	*(2.41)*	0.54	0.88	*(2.36)*
UK	3.26	*(8.05)*	2.60	3.03	*(7.96)*	1.57	*(5.02)*	1.33	1.26	*(4.09)*
USA	2.51	*(5.67)*	1.51	2.50	*(7.26)*	0.79	*(2.45)*	0.70	0.80	*(3.05)*
Panel B: Emerging markets										
Argentina	1.62	*(2.43)*	0.66	1.72	*(2.52)*	1.83	*(3.57)*	0.76	2.03	*(3.69)*
Brazil	1.12	*(1.85)*	0.49	1.13	*(1.84)*	0.60	*(1.56)*	0.40	0.61	*(1.68)*
Chile	1.22	*(3.05)*	0.75	1.27	*(3.37)*	0.29	*(1.16)*	0.28	0.26	*(1.07)*
China	2.36	*(10.00)*	2.41	2.14	*(9.08)*	1.42	*(4.58)*	1.15	1.22	*(4.40)*
Colombia	2.57	*(4.29)*	0.97	2.52	*(3.95)*	0.00	*(0.00)*	0.00	−0.27	*(−0.66)*
India	3.28	*(7.51)*	2.28	3.20	*(7.52)*	1.28	*(3.47)*	0.84	1.37	*(4.11)*
Indonesia	2.25	*(5.80)*	1.30	2.12	*(5.35)*	1.15	*(2.73)*	0.78	1.11	*(2.85)*
Korea	3.31	*(7.02)*	2.27	3.19	*(6.31)*	1.85	*(4.72)*	1.54	1.84	*(4.54)*
Kuwait	1.30	*(2.26)*	0.82	1.30	*(1.83)*	0.92	*(2.91)*	1.07	1.11	*(2.90)*
Malaysia	2.79	*(9.31)*	2.48	2.89	*(10.43)*	1.73	*(7.56)*	1.68	1.69	*(6.54)*
Mexico	0.90	*(2.16)*	0.52	0.90	*(2.16)*	0.67	*(1.81)*	0.40	0.63	*(1.73)*
Pakistan	1.36	*(1.86)*	0.60	1.39	*(1.90)*	0.78	*(2.27)*	0.52	0.87	*(2.39)*
Peru	0.43	*(0.52)*	0.16	0.51	*(0.58)*	1.20	*(2.16)*	0.68	1.07	*(1.94)*
Philippines	1.95	*(4.69)*	1.13	2.11	*(5.51)*	0.89	*(2.47)*	0.56	0.89	*(2.24)*
Poland	2.44	*(5.57)*	1.60	2.49	*(5.85)*	1.93	*(3.96)*	1.15	1.97	*(4.40)*
Qatar	1.92	*(2.57)*	1.01	2.44	*(3.12)*	−0.20	*(−0.19)*	−0.12	−0.79	*(−0.84)*
Russia	1.02	*(0.67)*	0.33	0.86	*(0.53)*	0.96	*(1.75)*	0.79	0.77	*(1.34)*
Saudi Arabia	1.27	*(2.42)*	1.03	1.18	*(1.87)*	0.22	*(0.39)*	0.19	−0.02	*(−0.04)*
South Africa	3.57	*(9.17)*	2.54	3.55	*(9.00)*	1.02	*(3.45)*	0.91	0.82	*(2.97)*
Taiwan	2.17	*(8.14)*	2.28	2.17	*(8.23)*	1.15	*(4.58)*	1.05	1.11	*(4.19)*
Thailand	2.02	*(5.24)*	1.38	2.00	*(4.88)*	1.62	*(5.55)*	1.30	1.43	*(4.88)*
Turkey	2.04	*(4.90)*	1.27	1.99	*(4.35)*	0.26	*(0.48)*	0.12	0.24	*(0.48)*
UAE	1.99	*(3.03)*	1.08	1.41	*(2.01)*	0.22	*(0.24)*	0.08	−0.10	*(−0.09)*
Panel C: International averages										
Global markets	2.39	*(5.39)*	1.44	2.37	*(5.38)*	1.06	*(3.01)*	0.77	0.99	*(2.80)*
Developed markets	2.83	*(6.23)*	1.61	2.80	*(6.31)*	1.17	*(3.35)*	0.83	1.11	*(3.14)*
Emerging markets	1.95	*(4.54)*	1.28	1.93	*(4.44)*	0.95	*(2.66)*	0.70	0.86	*(2.47)*

The table reports the mean monthly return (μ) and six-factor model alphas (α) on the long-short strategies based on the forecast combination model. For each country, each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The strategies are implemented separately in small and big firms, defined as those above and below the monthly median market capitalization. The portfolios are value-weighted and rebalanced monthly. μ *and* α are reported in percentage terms. The numbers in parentheses are Newey and West's (1987) adjusted t-statistics. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

4.4. Practical investor perspective

Besides understanding the role of small firms, we are also interested in two further practical aspects of machine learning strategies: robustness to transaction costs and reliance on very recent information. Both issues affect the eventual tradability and portfolio implementation of machine learning signals.

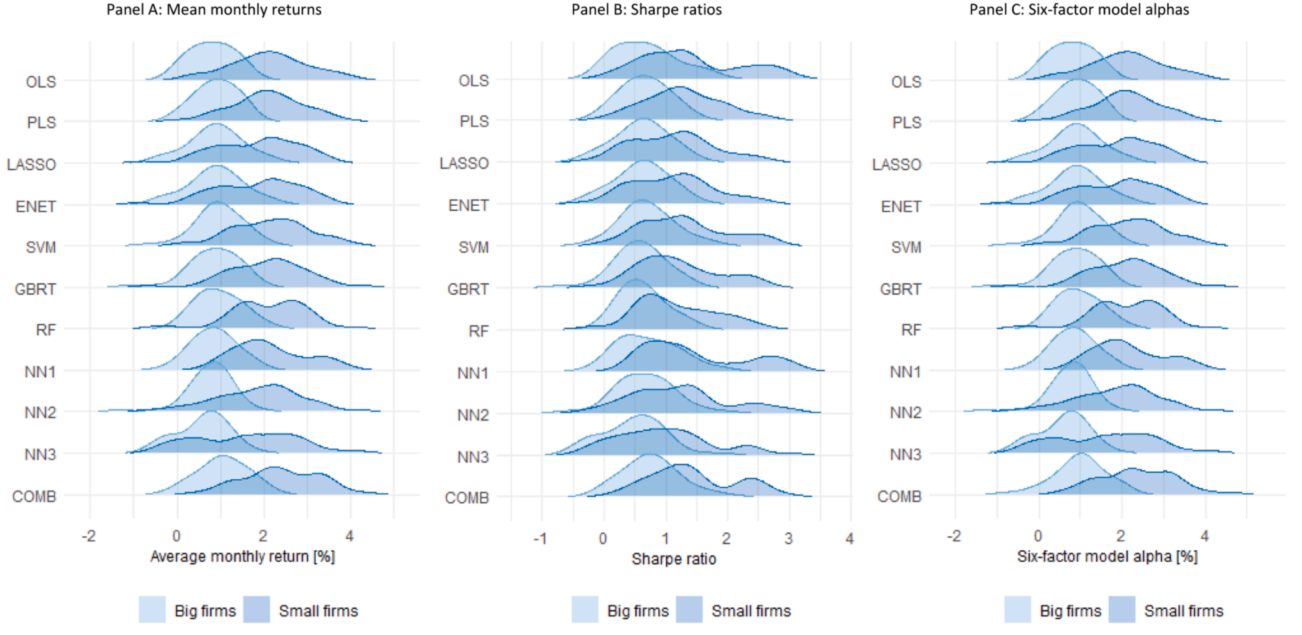

Fig. 7. Performance Distributions for Machine Learning Strategies in Small and Big Firms
The figure exhibits Gaussian kernel density plots for the mean monthly returns (Panel A), annualized Sharpe ratios (Panel B), and six-factor model alphas (Panel C) of different machine learning strategies (see Section 2.3) across the 46 stock markets that are covered in this study. For each country, each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The strategies are implemented separately in big (light-blue shades) and small (dark-blue shades) firms, defined as those above and below the monthly median market capitalization. The portfolios are value-weighted and rebalanced monthly. The total study period is from January 1991 to December 2020; the testing period starts in January 2001. The returns and alphas are reported in percentage terms.

4.4.1. Portfolio turnover and trading costs

Extant machine learning literature frequently accentuates the problem of high portfolio turnover (Gu et al., 2020; Leippold et al., 2022; Azevedo et al., 2022). Furthermore, Avramov et al. (2023) demonstrate that complex machine learning models concentrate on difficult-to-arbitrage stocks—where trading is costly. This can lead to substantial transaction costs, impeding the practical implementation of machine learning strategies.

Table 7 shows the average portfolio turnover and breakeven transaction costs for different machine learning models across countries in our sample.[15] We estimate the portfolio turnover for month t (PT_t) as in Bollerslev et al. (2018) and Koijen et al. (2018). Specifically, we calculate the average portfolio share that needs to be replaced each month:

$$PT_t = \frac{1}{2}\sum_{i=1}^{n}\left|w_{i,t-1}\times\left(1+r_{i,t}\right)-w_{i,t}\right|, \tag{7}$$

where $w_{i,t-1}$ and $w_{i,t}$ are the weights of stock i in two consecutive months, and $r_{i,t}$ denotes the stock return. To avoid double-counting the sells and buys, we calculate a one-sided (rather than two-sided) measure.[16] Finally, with the turnover estimation at hand, we compute the breakeven trading costs as the average portfolio return divided by the average turnover.

The portfolio turnover on machine learning strategies is generally high (Table 7, Panel A). Globally, the average values typically range between 80% and 140%. The lowest turnover characterizes simple strategies, such as OLS and PLS, while the complex neural networks exhibit the highest scores. The combination strategies are in the middle, with an average global turnover of 110%.

Table 7, Panel B, shows the breakeven trading costs. Interestingly, the simple strategies (OLS, PLS) seem more resilient to transaction costs than the complex ones (e.g., NN1, NN2, NN3). Overall, the breakeven trading costs are higher in developed markets than in emerging ones, benefiting from higher average returns. The top-performing strategy in this segment is COMB, deepening its superiority demonstrated in earlier tests. The average breakeven trading costs on COMB in the developed markets equal 2.04%.

4.4.2. Reliance on recent information

Earlier empirical evidence has suggested an essential role of recent data in forming machine learning forecasts. The most important return predictor in the U.S. market is the short-term reversal (Gu et al., 2020), which builds on last month's data. In China, the two variables contributing the most are volume volatility and the number of zero-trading days—both derived from the last month's daily data (Leippold et al., 2022). However, using this short-lived information in trading practice may be challenging or even unfeasible. For example, the short-term reversal anomaly relies on the closing price on the day, which is simultaneously assumed as the moment of portfolio formation for further tests.

To some extent, also our results reveal the importance of recent information. Fig. 3 shows that the short-term reversal is the second most prominent variable according to the VI ranking. Therefore, to assess the short-lived nature of machine learning signals, we reproduce our essential portfolio tests with an additional one-month skip period. To be precise, we lag all stock characteristics by one month. In consequence, we re-estimate all the models for each country by using information from up to t-1 to predict returns in $t+1$.

Table 8 reports the average performance of the strategies based on this alternative implementation across all the markets in the sample.[17] Foreseeably, the additional skip period has visibly impaired the models' effectiveness. Globally—across countries and firm types—the Sharpe ratios of long-short value-weighted COMB portfolios have declined by 18.9% (Table 8, Panel A). Nonetheless, the drop in performance was not equal everywhere. The Sharpe ratios decreased particularly in small stocks (on average 23.2%) rather than in big stocks (on average 15.5%).

When the individual models are considered (Table 8, Panel B), the negative impact of the additional skip period grows along with a model's complexity. For example, while the Sharpe ratios on LASSO and ENET were only moderately affected (on average −14.9% and −15.4%, respectively), the risk-adjusted performance of NN3 deteriorated by 29.8%. As a result—once the skip period is introduced—NN3 fares worst among all individual strategies according to all performance measures. In summary, while the very recent information is not critical to the success of machine learning strategies, it matters a lot—particularly for complex models.

5. International variation in machine learning returns

Our global analysis reveals a noticeable heterogeneity in machine learning profits around the world. Although the tested strategies are typically profitable everywhere, the magnitude of abnormal returns differs substantially across countries. Hence, this section explores the sources of this international variation. We embark on two main methods: market-level regressions and country sorts.

5.1. Baseline methodology

In this exercise, we are interested in scrutinizing the differences in machine learning profits *between* individual markets. Hence, to isolate the cross-sectional variation, we follow the methodological approach of Chui et al. (2010)—which was originally designed to

[15] Whereas the main paper for brevity presents only the international averages, Tables A18 and A19 in the Online Appendix provide detailed statistics for individual markets.

[16] Importantly, the average turnover calculated this way may still exceed 100% in the case of long-short portfolios.

[17] Detailed statistics for specifications in individual markets are available upon request.

Table 7
Portfolio turnover and breakeven trading costs of machine learning strategies.

	OLS	PLS	LASSO	ENET	SVM	GBRT	RF	NN1	NN2	NN3	COMB
Panel A: Portfolio turnover											
Global markets	87.9	79.6	131.9	131.9	130.8	131.3	126.1	140.1	136.2	133.9	110.5
Developed markets	94.9	85.8	138.5	138.5	139.1	134.4	128.0	145.5	140.5	139.7	88.2
Emerging markets	81.0	73.3	125.2	125.3	122.5	128.3	124.2	134.6	132.0	128.2	132.8
Panel B: Breakeven transaction costs											
Global markets	1.38	1.60	0.92	0.92	1.05	1.01	1.14	0.87	0.80	0.67	1.52
Developed markets	1.46	1.73	1.07	1.08	1.17	1.17	1.31	1.01	1.04	1.03	2.04
Emerging markets	1.30	1.47	0.76	0.76	0.93	0.84	0.97	0.73	0.57	0.30	1.00

The table presents the average portfolio turnover and breakeven transaction costs for different machine strategies (see Section 2.3) across the 46 global markets, 23 developed markets, and 23 emerging markets. For each country, each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The portfolios are value-weighted and rebalanced monthly. The portfolio turnover (Panel A) is calculated following Eq. (7) as the average portfolio share replaced each month. The average breakeven trading costs (Panel B) are computed as the average portfolio return divided by the average turnover. The underlying values for individual countries are available in Tables A18 and A19 in the Internet Appendix. The total study period is from January 1991 to December 2020; the testing period starts in January 2001. All values are reported in percentage terms.

Table 8
Performance of the machine learning strategies with a one-month skip period.

	Mean return (μ)			Sharpe ratio (SR)			Alpha (α)		
	Average	*t*-stat	%decline	Average	*t*-stat	%decline	Average	*t*-stat	%decline
Panel A: Combination strategy (COMB)									
All markets	1.30	*(13.43)*	−14.3	0.89	*(12.41)*	−18.9	1.28	*(13.51)*	−12.0
Developed markets	1.49	*(11.26)*	−13.8	0.90	*(9.46)*	−23.4	1.45	*(10.74)*	−12.5
Emerging markets	1.12	*(9.05)*	−14.9	0.87	*(8.98)*	−13.6	1.11	*(9.57)*	−11.4
Small firms	1.96	*(20.81)*	−18.2	1.11	*(15.45)*	−23.2	1.88	*(19.50)*	−20.8
Big firms	0.91	*(12.68)*	−14.0	0.65	*(11.25)*	−15.5	0.89	*(12.49)*	−9.9
Panel B: Individual strategies									
OLS	1.03	*(10.56)*	−13.8	0.80	*(10.16)*	−19.1	1.01	*(10.37)*	−14.0
PLS	1.12	*(11.97)*	−10.8	0.79	*(11.76)*	−14.2	1.10	*(11.74)*	−7.9
LASSO	1.11	*(11.78)*	−10.4	0.71	*(11.63)*	−14.9	1.07	*(11.09)*	−10.2
ENET	1.10	*(11.69)*	−11.0	0.71	*(11.50)*	−15.4	1.07	*(11.00)*	−10.9
SVM	1.15	*(10.38)*	−18.4	0.76	*(10.30)*	−23.3	1.09	*(9.38)*	−18.6
GBRT	1.13	*(10.13)*	−15.0	0.70	*(10.81)*	−18.7	1.07	*(9.53)*	−15.9
RF	1.19	*(11.12)*	−17.1	0.71	*(10.92)*	−20.4	1.13	*(10.37)*	−17.4
NN1	1.02	*(10.68)*	−16.5	0.81	*(10.11)*	−22.0	1.01	*(10.63)*	−14.0
NN2	0.96	*(10.96)*	−14.0	0.75	*(10.65)*	−21.0	0.95	*(10.62)*	−10.8
NN3	0.69	*(6.26)*	−25.2	0.51	*(6.01)*	−29.8	0.69	*(6.27)*	−23.5

The table presents the average performance of machine learning strategies (see Section 2.3) across international markets with an additional skip period. All the models are trained using information lagged by one month, i.e., from month t-1, to predict returns in month $t+1$. For each country each month, we form long-short portfolios that buy (sell) the quintile of stocks with the highest (lowest) predicted return. The portfolios are value-weighted and rebalanced monthly. Panel A reports the results for the COMB strategy in different configurations – across all 46 global markets, 23 developed markets, and 23 emerging markets, as well as in big and small firms only across 46 global markets. The big and small firms are defined as in Table 6, i.e., as those below and above median market capitalization in a country in a given month. Panel B reports the average results across the 46 global markets for individual machine learning strategies. We report the mean monthly returns (μ), annualized Sharpe ratios (SR), and alphas (α) from the six-factor model of Fama and French (2018)—respectively. The numbers in parentheses are bootstrap *t*-statistics for cross-country averages.*% decline* represents a percentage drop in performance relative to the baseline methodological approach without the additional skip period. The mean returns, alphas, and*%decline* are expressed in percentage terms. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

study the heterogeneity in momentum profits around the world. To be specific, we regress the machine learning strategy returns on various country characteristics from asset pricing literature:

$$ML_{j,t} = \gamma_0 + \gamma_1 F_j + \gamma_2 M_{j,t} + \varepsilon_{j,t}, \tag{8}$$

where $ML_{j,t}$ denotes the return on the long-short quintile machine learning strategy in market j in month t, F_j is the vector of time-invariant country characteristics (such as cultural traits); $M_{j,t}$ is the set of time-varying explanatory variables (such as idiosyncratic risk) that are updated monthly; γ_0, γ_1, and γ_2 are the estimated regression coefficients; and $\varepsilon_{j,t}$ indicates the error term. We estimate Eq. (8) following the Fama-Macbeth (1973) procedure, with the *t*-values calculated using Newey and West's (1987) correction.

Our analyses involve an array of potential explanatory variables, as well as regression specifications. Building on Eq. (8), we begin by running univariate regressions considering the role of individual predictors. Next, as in Chui et al. (2010), we continue with multivariate tests within different categories of variables. Last, we run a *comprehensive* regression specification that incorporates all

covariates that prove significant in the first two steps. To mitigate the risk of spurious findings and so-called Type-1 errors, we embark on a multiple hypotheses framework. Concretely, we verify the statistical significance of the coefficients using the Bonferroni adjustment in the tests.[18]

In the baseline tests, we apply the regressions to the value- and equal-weighted long-short strategies formed using the COMB method, as it most comprehensively summarizes individual prediction techniques. In the further robustness checks, we extend our findings to other prediction models.

5.2. Market characteristics

We employ a comprehensive spectrum of market features that may determine the returns on machine learning strategies. These characteristics could be classified into four broad groups: cultural traits, limits to arbitrage, market development, and sample structure. In this section, we provide only their brief overview; the details on the data sources, calculation methods, and statistical properties are available in Tables A20 and A21 in the Internet Appendix.

First, substantial evidence documents the influence of certain cultural traits on stock mispricing. In particular, the level of individualism (*IND*) and long-term orientation (*LTO*) may interplay with different behavioral biases and augment subjective probability distortion. Consequently, they tend to affect the magnitude of various anomalies associated with past returns and their distributions (Chui et al., 2010; Cheon and Lee, 2018; Docherty and Hurst, 2018; Gao et al., 2018; Hollstein and Sejdiu, 2020). Because these categories of signals constitute an important input into our machine learning models, we control for these two dimensions of national culture in our regressions.

Second, we consider several indicators of limits to arbitrage. The machine learning models heavily rely on return predicting variables that, at least, may partly manifest stock mispricing. This is because the mispricing tends to be amplified in the market segments characterized by heavy arbitrage constraints, which either hinder or delay the activities of sophisticated investors (Shleifer and Vishny, 1997). The limits to arbitrage are typically proxied by idiosyncratic risk (e.g., Ali et al., 2003; Brav et al., 2010; McLean, 2010; Lam and Wei, 2011) and various impediments to trading efficiency; such as short selling unavailability (Chu et al., 2020; Stambaugh et al., 2012), low liquidity (Sadka and Sherbina, 2007; Chordia et al., 2008; Lam and Wei, 2011), or even simple firm size (Azevedo and Müller, 2020). The arbitrage effectiveness may also be indirectly affected by interest rate levels via the funding channel (Brunnermeier and Pedersen, 2009; Jacobs, 2015; Bessembinder et al., 2021). Building on this background, we weigh in five proxies for arbitrage constraints: local short-term interest rate (*INT*), average stock illiquidity measured with the Amihud's ratio (*ILLIQ*), average idiosyncratic risk (*IRISK*), average firm size (*SIZE*), and short-sale permission (*SHORT*).

Second, we look at several indicators of arbitrage limits. Machine learning models rely heavily on return-predicting variables that may, to some extent, reflect stock mispricing. This reliance is rooted in the notion that mispricing is often exacerbated in market segments that face substantial arbitrage constraints. These constraints can either inhibit or delay the activities of sophisticated investors (Shleifer and Vishny, 1997), allowing mispricing to persist longer.

In the asset pricing literature, one of the most common proxies for arbitrage constraints is the idiosyncratic risk (Ali et al., 2003; Brav et al., 2010; McLean, 2010; Lam and Wei, 2011). Elevated levels of this risk introduce unpredictability into arbitrage activity, which deters arbitrageurs who typically seek lower-risk discrepancies in asset prices. Assets with high idiosyncratic risk, unique to a particular firm and influenced by firm-specific factors, are more difficult to hedge. Furthermore, idiosyncratic risk is often associated with higher transaction costs. Such an environment makes exploiting potential mispricing more challenging and increases the risk of arbitrage strategies. In addition, markets characterized by high idiosyncratic risk often exhibit greater uncertainty about company fundamentals, which poses an additional challenge to arbitrageurs and further limits arbitrage activity.

This confluence of factors affects the informational efficiency of markets. In environments with high idiosyncratic risk, the speed and accuracy with which information is incorporated into asset prices may be compromised, leading to lower levels of informational efficiency. Thus, idiosyncratic risk affects not only the feasibility of arbitrage but also the predictive power of machine learning models under different market conditions.

Additional barriers to efficient trading, such as the unavailability of short selling (Chu et al., 2020; Stambaugh et al., 2012), low liquidity (Sadka and Sherbina, 2007; Chordia et al., 2008; Lam and Wei, 2011), or even the size of the firm (Azevedo and Müller, 2020), can also increase the limitations of arbitrage. For example, if a market has poor liquidity or lacks the infrastructure for short selling, it reduces the ability of investors to quickly correct mispricing, potentially increasing the duration and intensity of price anomalies.

In addition, the effectiveness of arbitrage may be indirectly affected by the level of interest rates through the funding channel (Brunnermeier and Pedersen, 2009; Jacobs, 2015; Bessembinder et al., 2021). A high-interest rate environment may increase the cost of borrowing and, thus, the cost of executing arbitrage strategies, creating an additional obstacle to arbitrage.

Taking all of these factors into account, we have chosen five proxies for arbitrage constraints in our analysis: the local short-term interest rate (*INT*), average stock illiquidity as measured by the Amihud ratio (*ILLIQ*), average idiosyncratic risk (*IRISK*), average firm size (*SIZE*), and short-selling permission (*SHORT*). This comprehensive approach allows us to capture the multifaceted nature of arbitrage constraints across markets.

The third category of variables concerns market development and efficiency. The common narrative suggests that return

[18] Given the 16 different explanatory variables considered in the regressions (see Section 5.2), the 5% significance level (two-side test) implies a Bonferroni-corrected *t*-statistic of 2.96.

predictability should be stronger within emerging markets. Developing financial markets are archetypally less efficient, and new information cannot be quickly incorporated into prices. This, in turn, provides fertile ground for the occurrence of mispricing—which is captured by numerous return predictors underlying machine learning strategies. Against this backdrop, the empirical evidence is mixed—with certain studies finding more prevalent mispricing in emerging markets (Batram and Grinblatt, 2021) and others finding no significant difference (Griffin et al., 2010; Jacobs, 2016). Our study employs several proxies of market development and efficiency from earlier studies (e.g., Watanabe et al., 2013; Jacobs, 2016; Cakici and Zaremba, 2022; Azevedo et al., 2022). The selection includes the return synchronicity measure (*SYNCH*) by Morck et al. (2000), the future earnings response coefficient (*FERC*) originating from Collins et al. (1994) and Durnev et al. (2003), and the binary developed market indicator (*DEV*) of Azevedo and Müller (2020).[19] Furthermore, as in Cakici and Zaremba (2022a, 2022b), we control for financial market openness (*OPEN*) measured with the Chinn-Ito index (Chin and Ito, 2006)—which correlates with market development and may facilitate mispricing elimination.[20]

Fourth, we examine the role of country sample structure characteristics—particularly the number of listed firms (*FIRMS*)—that may affect strategy returns. This focus stems from the inherent relationship between the size of a market, in terms of its listed firms, and the effectiveness of predictive models. Machine learning algorithms tend to thrive when provided with larger, richer datasets, enabling the construction of more robust predictive models. This fact underscores the critical role played by the number of listed firms in a market. A more significant number of firms directly translates into a larger volume of data, which increases the training effectiveness of machine learning models and, in turn, their predictive power.

The size of the stock universe, as determined by the number of listed firms (*FIRMS*), is likely to affect machine learning returns through a couple of critical channels. First, a bigger number of stocks implies more observations, offering a more comprehensive dataset for training more reliable predictive models. Second, as suggested by Bessembinder et al. (2021), return predictability is correlated with the diversity of listed firms, leading to greater market complexity. This relationship stems from the diversity of firm characteristics within larger markets and the subsequent potential for anomaly detection and exploitation by machine learning algorithms.

In addition to the sheer volume of data, the quality and diversity of the information encapsulated in the stock characteristics used as model inputs significantly impact predictive accuracy. Therefore, we consider the number of available features (*FTRS*) and the length of available data time series (*TSL*). More extended time series and an increased number of variables invariably increase the potential for superior predictive models, further emphasizing the importance of a larger market size with a large number of listed companies.

To control for these effects, we employ two measures proposed by Bessembinder et al. (2021): diversity of firm characteristics (*DIV*), which reflects the breadth of unique firm profiles within the market, and economic complexity (*ECOMP*), which is calculated as the Herfindahl-Hirschman index based on industry sales data. These metrics allow us to capture the complex dynamics of large markets and shed light on the role of market size in determining the effectiveness of machine learning strategies.

5.3. Regression results

Table 9 reports the results of cross-sectional regressions, with Panels A and B pertaining to value- and equal-weighted strategies—respectively. Columns (1) in both panels display the outcomes of univariate tests. Columns (2) to (5) concern multivariate regressions that are run within different groups of variables (cultural traits, limits to arbitrage, market development, and sample structure). Finally, Column (6) reports the result of *comprehensive tests* that incorporate only those variables that are significant across all univariate and multivariate regressions in Columns (1) to (5).

The univariate tests reveal that several factors may indeed matter for machine learning profitability around the world. Though the cultural traits do not reveal any reliable impact, the limits to arbitrage certainly play some role. Machine learning returns are boosted by high *IRISK*—which belongs to the most popular arbitrage constraint indicators. The essential role of *IRISK* corroborates the earlier findings of Avramov et al. (2023), who argue that machine learning strategies extract profitability from difficult-to-arbitrage market segments. Besides the idiosyncratic risk, the return predictability seems to be stronger in small firms; however, the statistical significance of this relationship is seen to be weaker. Furthermore, small companies are usually riskier; therefore, the two variables may capture the same economic phenomenon. Importantly, when considered jointly, *IRISK* prevails—fully subsuming *SIZE*.

The market development variables do not uncover a significant association with the machine learning profit. Perhaps the only exception of the positive impact of *DEV* in certain specifications complies with the earlier conclusions from Tables 2 and 3.

Finally, the overview of the sample structure variables discloses that the number of firms in the market (*FIRMS*) strongly augments the machine learning profits. As we have already noted, a larger equity universe enables better model training and strengthens return predictability by stock market factors. Furthermore, the diversity of firm characteristics (*DIV*) plays a partial role in increasing strategy

[19] Market development is also associated with accounting quality, which may also play a role in our findings. We observe weaker predictability in developing markets, where the accounting standard tends to be less accurate. Therefore, in an unreported analysis, we also experiment with several proxies for accounting quality, such as the loss avoidance ratio (Burghstahler & Dichev, 1997; Leuz et al., 2003), profit decline avoidance ratio (Burghstahler & Dichev, 1997), and accuals ratio (Dechow et al., 1995; Sloan, 1996; Leuz et al. 2004). We find no consistent impact of accounting quality. This is also consistent with our variable importance analysis, which generally emphasizes the market-based variables relative to accounting ones.

[20] In preliminary tests, we also consider the aggregate measure of stock market importance (*MKT*) from Watanabe et al. (2013). Since this indicator largely relies on the number of companies in the market, it correlates strongly with our "number of firms" variable (*FIRMS*). Importantly, *FIRMS* subsumes *MKT* in joint tests—so we limit our considerations to *FIRMS*.

Table 9
Determinants of the machine learning returns.

Panel A: Value-weighted portfolios						
	(1) Univariate tests	(2) Cultural traits	(3) Limits to arbitrage	(4) Market development	(5) Sample structure	(6) Compreh. test
IDV	0.070 (1.50)	0.077 (1.55)				
LTO	0.009 (0.31)	0.034 (1.24)				
INT	0.513 (1.80)		0.237 (0.68)			
ILLIQ	0.014 (2.06)		0.004 (0.60)			
IRISK	**0.599 (6.42)**		**0.489 (4.33)**			**0.501 (4.97)**
SIZE	**−2.799 (−2.99)**		−1.437 (−1.25)			
SHORT	−0.135 (−0.06)		1.887 (1.05)			
SYNCH	−0.025 (−1.52)			−0.022 (−1.36)		
FERC	−0.578 (−0.44)			−0.954 (−0.72)		
DEV	1.799 (0.85)			6.961 (2.42)		
OPEN	−2.265 (−1.80)			−3.718 (−2.49)		
FTRS	0.042 (1.24)				−0.103 (−2.06)	
TSL	0.008 (0.14)				0.006 (0.10)	
DIV	0.050 (2.87)				0.035 (1.82)	
FIRMS	**2.762 (7.80)**				**2.542 (4.10)**	**1.601 (3.96)**
ECOMP	−2.710 (−0.24)				6.594 (0.53)	
#Obs.		8976	6956	8412	9096	9096
$\overline{R^2}$		0.0222	0.0392	0.0199	0.0509	0.0258
Panel B: Equal-weighted portfolios						
	(1) Univariate tests	(2) Cultural traits	(3) Limits to arbitrage	(4) Market development	(5) Sample structure	(6) Compreh. test
IDV	0.057 (1.36)	0.070 (1.58)				
LTO	0.053 (1.90)	0.068 (2.34)				
INT	0.037 (0.12)		−0.138 (−0.34)			
ILLIQ	0.010 (1.95)		0.000 (0.02)			
IRISK	**0.458 (6.34)**		**0.401 (3.77)**			**0.367 (4.42)**
SIZE	−1.670 (−1.89)		0.147 (0.12)			
SHORT	−1.482 (−0.69)		−2.536 (−1.15)			
SYNCH	−0.020 (−1.43)			−0.015 (−1.09)		
FERC	0.509 (0.46)			−1.177 (−1.03)		
DEV	4.141 (2.13)			**9.352 (3.39)**		
OPEN	−0.481 (−0.39)			−2.847 (−1.85)		
FTRS	0.031 (1.06)				**−0.161 (−3.54)**	
AGE	0.044 (0.89)				0.039 (0.62)	
DIV	**0.041 (3.30)**				0.032 (2.16)	
FIRMS	**2.345 (5.68)**				**2.357 (3.24)**	**1.544 (3.36)**
ECOMP	−8.654 (−1.21)				3.206 (0.33)	
#Obs.		8976	6956	8412	9096	9096
		0.0267	0.0563	0.0192	0.0505	0.0177

The table reports the average slope coefficients of cross-sectional regressions of monthly returns on single-country machine learning strategies based on the combination model (COMB) on potential drivers of machine learning strategy returns. The dependent variable is the return on the single-market long-short portfolio buying (selling) a quintile of stocks with the highest (lowest) prediction from the COMB model. The strategies are applied in each of the 46 stock markets in our sample. The portfolios are either value-weighted (Panel A) or equal-weighted (Panel B) and rebalanced monthly. The explanatory variables fall into four categories: a) cultural traits: individualism (*IND*) and long-term orientation (*LTO*); b) limits to arbitrage: local interest rate (*INT*), illiquidity (*ILLIQ*), idiosyncratic risk (*IRISK*), average firm size (*SIZE*), and short-selling permission (*SHORT*); c) market development: return synchronicity (*SYNCH*), future earnings response coefficient (*FERC*), developed market indicator (*DEV*), and financial openness (*OPEN*); and d) sample structure: number of features (*FTRS*), time-series length (*TSL*), diversity in stock characteristics (*DIV*), number of publicly-listed firms (*FIRMS*), and economic complexity (*ECOMP*). Table A18 in the Internet Appendix details the market characteristics. Column (1) displays the coefficients of univariate regressions; columns (2) to (5) present the coefficients from multivariate regressions conducted within different categories of variables; finally, column (6) presents the coefficients from the comprehensive model. The numbers in parentheses are Newey and West's (1987) adjusted *t*-statistics. *#Obs.* is the number of monthly observations. $\overline{R^2}$ is the average cross-sectional adjusted R-squared coefficient. The total study period is from January 1991 to December 2020; the testing period starts in January 2001. All coefficients are multiplied by 1000 except for *IRISK*, *SYNCH*, and *DIV*. The underline font indicates coefficients that pass statistical significance at the 5% level after the Bonferroni adjustment for multiple hypotheses.

returns. Both observations correspond with the arguments of Bessembinder et al. (2021), who document their central role in shaping return predictability within the U.S. market.

To sum up, two variables in Table 9 demonstrate prominent importance: *IRISK* and *FIRMS*. Both are highly significant across all specifications; this includes univariate and multivariate tests of both equal- and value-weighted portfolios—easily exceeding the Bonferroni-corrected significance threshold. Considering them together in a joint regression (Column [6]) confirms that they both matter for the profitability of machine learning strategies. To conclude, the international heterogeneity of machine learning returns is primarily driven by the number of listed firms in the market and local limits to arbitrage proxied with idiosyncratic volatility. Taken

together, they explain on average 2.58% (1.77%) of the cross-sectional variation in value-weighted (equal-weighted) portfolio returns.

Finally, we are also interested in the relationship between *IRISK* and *FIRMS* and other market characteristics considered in this study. For example, our earlier evidence suggested that return predictability is stronger in developed markets. Meanwhile, the developed markets commonly list more companies, so the two variables may simply capture the same economic phenomena. Furthermore, large markets may be home to numerous small companies whose limited correlation with the main market index pushes the average idiosyncratic risk upwards.[21]

Another point is the importance of the firm size. Smaller firms exhibit stronger return predictability, and Table 9 shows that the average firm size may play some role in machine learning profitability. However, small firms typically have higher idiosyncratic volatility, so *SIZE* and *IRISK* tend to be associated with each other.

To ascertain that other variables do not subsume *IRISK* and *FIRMS*, we control for them jointly in supplementary regressions reported in Table A22 in the Internet Appendix. The results confirm that *IRISK* and *FIRMS* remain significant in all specifications, surviving the impact of all control variables. On the other hand, noteworthily, they explain the impact of nearly all other variables, including *DEV*. In other words, we observe superior returns in developed markets because they are populated by many firms, many of which have small sizes and high idiosyncratic risk. Once we control the number of firms and their average idiosyncratic volatility, market development no longer plays any role.[22]

5.4. Country sorts

To further illustrate how *FIRMS* and *IRISK* shape the international variation in machine learning returns, we now carry on with country sorts. In this exercise, we group different markets together based on *FIRMS* and *IRISK* to observe the cross-sectional differences in average machine learning returns in countries with different numbers of firms and levels of idiosyncratic risk.

As in Table 9, in this baseline approach, we concentrate on the *COMB* strategy. We run two types of country sorts: univariate and bivariate. The univariate sorts serve as an acid test for monotonic patterns in country-specific machine learning returns. Each month, we rank countries on *FIRMS* or *IRISK* and then sort them into terciles. In addition, we calculate differential returns on paper strategies assuming a long (short) position in the machine learning strategies in countries with the highest (lowest) *FIRMS* or *IRISK*.

In turn, bivariate sorts aim to ascertain that both variables contain both unique and independent information about future returns. We want to ensure that *FIRMS* and *IRISK* capture separate economic phenomena and remain significant after controlling for each other. Hence—for example—in order to explore the incremental information of *FIRMS*, we run the following procedure. First, we rank countries on *IRISK* and group them into terciles. Second, within each of the *IRISK* sets, we sort countries on *FIRMS*—effectively producing 9 [$=3 \times 3$] double-sorted portfolios. Third, we calculate average returns on markets with a consistent level of *FIRMS* across different *IRISK* terciles. Similarly, as performed in earlier tests, we also calculate spread portfolios that assume long (short) positions in top (bottom) *FIRMS* countries and evaluate them with a global six-factor model of Fama and French (2018). Finally, we run the analogous procedure to assess the incremental role of the *IRISK* subject to the impact of *FIRMS*.

Table 10 reports the results of country groupings. Both sorts on *FIRMS* (Panel A) and *IRISK* (Panel B) reveal evident patterns in the cross-section of country-specific *COMB* strategies. Observe first: the univariate sorts. For the value-weighted (equal-weighted) portfolios, the machine learning strategies in the tercile of countries with a higher number of listed firms outperform those in the markets with the lowest number of firms by 0.98% (0.84%) per month. In turn, the differential return on the value-weighted (equal-weighted) country portfolios formed on *IRISK* equals 0.92% (0.80%). The differences are highly significant in both cases and cannot be explained by the six-factor model.

Fig. 8 illustrates the cumulative returns on country terciles sorted on *FIRMS* and *IRISK*. Notably, the influence of the two variables is remarkably stable through time. There is very little variation in their influence on machine learning profits. The strategies implemented in markets with many firms and high idiosyncratic risk systematically beat their counterparts in countries with fewer firms and lower risk.

Leaving aside the univariate sorts, the two-way sorts in Table 10 confirm that the number of firms and idiosyncratic risk are independent drives of machine learning returns. The two variables, *FIRMS* and *IRISK*, generate clear cross-sectional return patterns—even after accounting for each other. Let us take the value-weighted portfolios as an example. The average differential return on sorts of *FIRMS* after controlling for *IRISK* is 0.53%; next, the average differential return from country sorts on *IRISK* subject to *FIRMS* is 0.54%. Both values are highly significant and cannot be attributed to common global risk factors.

Interestingly, a closer look at the influence of *IRISK* in different firm terciles reveals its uneven influence. The idiosyncratic risk particularly matters in markets populated by many companies. Hence, in the top *FIRMS* subset, the sorting into terciles generates a differential return of 0.77% per month. On the other hand, the average monthly return spread between the top and bottom *IRISK* countries in the low *FIRMS* markets is only 0.23%.

[21] The average firm-level idiosyncratic volatility in our sample is slightly higher in developed markets than in emerging ones. The average *IRISK* in these two groups equals 0.0293 and 0.0266, respectively. In this regard, the patterns in idiosyncratic derived from stock-level data differ from their counterparts in country-level data; in the latter case, emerging markets typically display higher non-diversifiable risk (Umutlu, 2010).

[22] One interesting observation from Table A22 is the negative coefficient on *FTRS*. In other words, once we control for the number of firms and their average idiosyncratic risk, increasing the number of available features tend to negatively affect the machine learning profitability. While this may seem surprising, recall from Fig. 3 that the critical variables are relatively simple, thus, commonly available in most countries. Therefore, including more variables may add little to return predictability, but increase the risk of overfitting.

Table 10
Number of firms, idiosyncratic risk, and machine learning profitability.

Panel A: Sorts on *FIRMS*										
	Panel A.1: Value-weighted portfolios					Panel A.2: Equal-weighted portfolios				
	All markets	Low IRISK	Medium IRISK	High IRISK	Average	All markets	Low IRISK	Medium IRISK	High IRISK	Average
Low FIRMS	1.18	1.11	1.28	1.79	1.39	1.74	1.62	1.72	2.17	1.84
	(6.77)	*(6.63)*	*(4.73)*	*(5.54)*	*(7.37)*	*(10.66)*	*(8.86)*	*(6.99)*	*(7.49)*	*(10.12)*
Medium FIRMS	1.77	1.20	1.64	2.41	1.75	2.08	1.73	2.10	2.79	2.21
	(8.63)	*(6.02)*	*(6.78)*	*(9.52)*	*(8.92)*	*(11.64)*	*(9.39)*	*(9.87)*	*(12.82)*	*(12.95)*
High FIRMS	2.15	1.57	1.73	2.46	1.92	2.58	2.00	2.16	2.83	2.33
	(11.18)	*(7.88)*	*(9.79)*	*(10.20)*	*(10.68)*	*(15.00)*	*(12.26)*	*(12.36)*	*(13.24)*	*(14.67)*
High-Low R	0.98	0.46	0.44	0.68	0.53	0.84	0.38	0.44	0.66	0.49
	(7.32)	*(2.12)*	*(1.84)*	*(2.60)*	*(4.05)*	*(6.72)*	*(1.92)*	*(2.02)*	*(2.64)*	*(3.98)*
High-Low α	0.76	0.45	0.56	0.32	0.44	0.69	0.41	0.54	0.39	0.45
	(7.73)	*(2.08)*	*(2.55)*	*(1.08)*	*(4.57)*	*(5.02)*	*(1.77)*	*(2.25)*	*(1.34)*	*(2.91)*
Panel B: Sorts on *IRISK*										
	Panel B.1: Value-weighted portfolios					Panel B.2: Equal-weighted portfolios				
	All markets	Low FIRMS	Medium FIRMS	High FIRMS	Average	All markets	Low FIRMS	Medium FIRMS	High FIRMS	Average
Low IRISK	1.30	1.07	1.51	1.83	1.47	1.79	1.39	1.89	2.32	1.87
	(8.77)	*(6.02)*	*(7.17)*	*(11.74)*	*(10.05)*	*(12.82)*	*(8.02)*	*(10.80)*	*(15.71)*	*(13.92)*
Medium IRISK	1.56	1.17	1.57	1.95	1.56	2.01	1.85	2.04	2.43	2.11
	(8.45)	*(5.63)*	*(5.79)*	*(8.13)*	*(8.32)*	*(12.24)*	*(8.87)*	*(8.40)*	*(11.51)*	*(12.25)*
High IRISIK	2.22	1.30	2.12	2.60	2.01	2.60	2.02	2.26	2.95	2.41
	(9.45)	*(4.32)*	*(8.58)*	*(10.60)*	*(8.82)*	*(12.82)*	*(7.98)*	*(10.08)*	*(13.48)*	*(12.38)*
High-Low R	0.92	0.23	0.62	0.77	0.54	0.80	0.63	0.37	0.63	0.54
	(5.73)	*(0.81)*	*(2.72)*	*(4.32)*	*(3.44)*	*(6.02)*	*(2.58)*	*(1.81)*	*(3.85)*	*(4.12)*
High-Low α	0.79	0.34	0.40	0.76	0.50	0.72	0.78	0.23	0.54	0.52
	(4.47)	*(1.08)*	*(1.71)*	*(4.31)*	*(3.24)*	*(4.75)*	*(2.82)*	*(1.18)*	*(3.15)*	*(3.51)*

The table reports the average returns on machine learning strategies in markets grouped by the number of publicly listed firms (*FIRMS*) and average idiosyncratic risk (*IRISK*). The single-country long-short portfolios buy (sell) a quintile of stocks with the highest (lowest) prediction from the COMB model. The strategies are applied in each of the 46 stock markets in our sample. The portfolios are value-weighted (Panels A.1 and B.1) or equal-weighted (Panel A.2 and B.2) and rebalanced monthly. Panel A concentrates on the sorts *FIRMS,* and Panel B focuses on the sorts on *IRISK*. The leftmost column in each panel ("All markets") presents the results of univariate sorts into terciles (*Low, Medium, High*) on *FIRMS* and *IRISK*. *High-Low R* is the average differential return in a portfolio that is long (short) in the *High* (*Low*) tercile; furthermore, *High-Low α* is the associated alpha from the six-factor model of Fama and French (2018). The right section of each panel reports the bivariate sorts on *FIRMS* and *IRISK*. In the first step, we group markets into terciles based on the control variable indicated in the top row. Next, within each subset of the control variable, we sort markets into terciles based on *FIRMS* or *IRISK*. *Average* is the average return on the market groups with a consistent level of the primary variable across different levels of the control variable. Returns and alphas are reported in percentage terms. The values in parentheses are Newey and West's (1987) adjusted *t*-statistics. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

To sum up, the country sorts confirm the number of firms in the market and average idiosyncratic risk are the key drivers of returns on machine learning strategies. To further validate these findings, we supplement our analysis with two further robustness checks. First, we modify the definitions of idiosyncratic risk. We shorten the estimation period from a year to one month. Furthermore, we use the median rather than the average market value. We also derive the idiosyncratic volatility from different asset pricing models: the three-factor model of Fama and French (1993) and the q-model of Hou et al. (2015). The results indicate that the role of *IRISK* does not hang on its definition (see Table A23 in the Internet Appendix for details). Second, rather than focusing on the forecast combination strategy, we apply the country sorts to individual machine learning models. The results of this experiment, summarized in Table A24 in the Internet Appendix, confirm that the impact of *FIRMS* and *IRISK* is not limited to any particular algorithm. The two significantly affect the profitability of all machine learning models that are considered in our study.

6. Concluding remarks

This study examines the application of machine learning strategies to predict stock returns in international markets. Using CRSP and Compustat data, we calculate 148 stock characteristics to run 11 machine-learning models in 46 countries. We explore their performance and scrutinize their drivers around the world.

An initial evaluation of the machine learning methods reveals substantial variation in their prediction performance across both models and countries. The OLS typically exhibits the lowest predictive R^2 coefficients, while COMB is the best performer. The return predictability is visibly stronger in developed markets than in emerging ones. Last, all methods can rank stocks quite successfully in accordance with ex-post realized returns. In other words, all machine learning models can be translated into successful portfolio sorts. The traditional R^2 scores do not prove to be the perfect measure of potential economic gains from machine learning.

The most crucial stock characteristics feeding the models belong to the traditionally popular categories, such as value, size, momentum, and reversal. The models are in relative agreement on the selection of key features; however, their exact contribution is ununiform. The most important variables comprise the ratio of current price to the maximum price over the last year, short-term

Fig. 8. Cumulative Machine Learning Returns in Different Market Groups

The figure plots the cumulative returns on machine learning strategies in countries grouped on the number of publicly listed firms (*FIRMS*) and average idiosyncratic risk (*IRISK*). First, we calculate single-country long-short portfolios that buy (sell) a quintile of stocks with the highest (lowest) COMB model forecast. The strategies are applied in each of the 46 markets in our sample. Next, we group the countries into terciles based on *FIRMS* (Panels A.1 and A.2) or *IRISK* (Panels B.1 and B.2). Finally, we additively cumulate the returns in each of the terciles (*Low, Mid, High*). The portfolios are value-weighted (Panels A.1 and B.1) or equal-weighted (Panel A.2 and B.2) and rebalanced monthly. Additionally, the gray line represents the cumulative excess return on the global market portfolio, represented by the MSCI ACWI Index. The values are in percentages. The total study period is from January 1991 to December 2020; the testing period starts in January 2001.

reversal, performance mispricing factor, earnings to price, age, market equity, book-to-market equity, and three-, six-, and nine-month price momentum.

All machine learning models can be forged into successful trading strategies. This even pertains to the commonly disregarded OLS, which performs at par with more complex algorithms. Again, the best strategy is forecast combination—highlighting the benefits of reducing forecast variance by merging multiple predictions. Globally, the long-short portfolio buying (selling) a value-weighted quintile of stocks with the highest (lowest) prediction from the COMB model earns 1.51% per month at an annualized Sharpe ratio of 1.49. Last, despite the overly robust performance, the machine learning strategies exhibit variation in profitability across different dimensions. In particular, the models work noticeably better in both developed markets and small firm segments.

Last, our study explores the sources of international differences in the efficiency of machine learning models. To unearth the drivers of heterogeneity, we investigate a range of market characteristics from different domains: cultural traits, limits to arbitrage, market development, and sample structure. We find the machine learning returns are particularly boosted by two features: the number of firms in the sample and idiosyncratic risk. The size of the equity universe not only allows for better training of the models but also strengthens the factor structure in returns. The idiosyncratic risk, in turn, is one of the most common gauges of limits to arbitrage—which tend to strengthen mispricing and return predictability. Both variables are robust across various tests and survive after accounting for each other, as well as within a multiple hypotheses testing framework.

Our findings bear direct practical implications. The machine learning models may serve as efficient tools to select stocks and form portfolios; however, their effectiveness is not uniform. The performance of machine learning strategies varies across many dimensions, depending on the model type, market segment, or country characteristics. Factors such as the number of available firms or local limits to arbitrage may decide on the model's successful implementation within equity markets.

Future studies on the topics discussed in our paper may further explore the impact of trading costs and implementation constraints on machine learning strategies. Avramov et al. (2023) demonstrated that machine learning predictions are subject to considerable economic restrictions, which may undermine potential economic gains. This issue becomes even more appalling internationally as moving capital across countries and currencies further impedes investment performance. International economic constraints may be a key factor determining practical gains from machine learning strategies.

Acknowledgments

We thank Vitor Azevedo, Turan Bali, Hendrik Bessembinder, David Blitz, Guillaume Coqueret, Guanhao (Gavin) Feng, Matthias Hanauer, Fabian Hollstein, Tobias Hoogteijling, Theis Ingerslev Jensen, Tobias Kalsbach, Herald Lohre, Alexandre Rubesam, Laurens Swinkels, Andea Tamoni, Pim van Vliet, and Guofu Zhou for helpful comments and suggestions, as well as seminar participants at Technical University of Munich, Robeco, Ono Academic College, American University of Sharjah, and American University in Dubai. Adam Zaremba acknowledges the support of the National Science Center of Poland [grants no. 2019/33/B/HS4/01021 and 2022/45/B/HS4/00451]. All errors are our own.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jedc.2023.104725.

References

Ali, A., Hwang, L.-S., Trombley, M.A., 2003. Arbitrage risk and the book-to-market anomaly. J. Financ. Econ. 69, 355–373.
Asness, C.S., Moskowitz, T.J., Pedersen, L.H., 2013. Value and momentum everywhere. J. Finance 68 (3), 929–985.
Avramov, D., Cheng, S., Metzker, L., 2023. Machine learning versus economic restrictions: evidence from stock return predictability. Manage. Sci. 69 (5), 2547–3155.
Avramov, D., Chordia, T., Goyal, A., 2006. Liquidity and autocorrelations in individual stock returns. J. Finance 61 (5), 2365–2394.
Azevedo, V., Hoegner, C., 2023. Enhancing stock market anomalies with machine learning. Rev. Quant. Financ. Account. 60, 195–230.
Azevedo, V., & Müller, S. (2020). Analyst recommendations and mispricing across the globe. Available at SSRN 3705141.
Azevedo, V., Kaiser, S., & Müller, S. (2022). Stock market anomalies and machine learning across the globe. Available at SSRN: https://ssrn.com/abstract=4071852 or doi: 10.2139/ssrn.4071852.
Bali, T., Goyal, A., Huang, D., Jiang, F., & Wen, Q. (2021). Predicting corporate bond returns: Merton meets machine learning. Georgetown McDonough School of Business Research Paper No. 3686164. Swiss Finance Institute Research Paper No. 20-110. Available at SSRN: https://ssrn.com/abstract=3686164.
Barber, B.M., De George, E.T., Lehavy, R., Trueman, B., 2013. The earnings announcement premium around the globe. J. Financ. Econ. 108 (1), 118–138.
Bartram, S.M., Grinblatt, M., 2021. Global market inefficiencies. J. Financ. Econ. 139 (1), 234–259.
Bates, J.M., Granger, C.W., 1969. The combination of forecasts. J. Oper. Res. Soc. 20 (4), 451–468.
Bessembinder, H., Burt, A.P., & Hrdlicka, C.M. (2021). Time series variation in the factor zoo. Available at SSRN: https://ssrn.com/abstract=3992041 or doi: 10.2139/ssrn.3992041.
Bianchi, D., Büchner, M., Tamoni, A., 2021. Bond risk premiums with machine learning. Rev. Financial Studies 34 (2), 1046–1089.
Bollerslev, T., Hood, B., Huss, J., Pedersen, L.H., 2018. Risk everywhere: modeling and managing volatility. Rev. Financial Studies 31 (7), 2729–2773.
Brav, A., Heaton, J.B., Li, S., 2010. The limits of the limits of arbitrage. Rev. Financ. 14 (1), 157–187.
Brunnermeier, M.K., Pedersen, L.H., 2009. Market liquidity and funding liquidity. Rev. Financial Studies 22 (6), 2201–2238.
Burgstahler, D., Dichev, I., 1997. Earnings management to avoid earnings decreases and losses. J. Account. Econ. 24 (1), 99–126.
Cakici, N., Zaremba, A., 2021a. Liquidity and the cross-section of international stock returns. J. Bank Financ. 127, 106123.
Cakici, N., & Zaremba, A. (2022). Empirical asset pricing via machine learning: the global edition. Available at SSRN 4028525.
Cakici, N., Zaremba, A., 2022b. Salience theory and the cross-section of stock returns: international and further evidence. J. Financ. Econ. 146 (2), 689–725.
Carhart, M.M., 1997. On persistence in mutual fund performance. J. Finance 52 (1), 57–82.

Chen, L., Pelger, M., Zhu, J., 2023. Deep learning in asset pricing. Manage. Sci. in press.
Cheon, Y.H., Lee, K.H., 2018. Maxing out globally: individualism, investor attention, and the cross section of expected stock returns. Manage. Sci. 64 (12), 5807–5831.
Chinn, M.D., Ito, H., 2006. What matters for financial development? Capital controls, institutions, and interactions. J. Dev. Econ. 81 (1), 163–192.
Chordia, T., Roll, R., Subrahmanyam, A., 2008. Liquidity and market efficiency. J. Financ. Econ. 87 (2), 249–268.
Chu, Y., Hirshleifer, D., Ma, L., 2020. The causal effect of limits to arbitrage on asset pricing anomalies. J. Finance 75 (5), 2631–2672.
Chui, A.C., Titman, S., Wei, K.C., 2010. Individualism and momentum around the world. J. Finance 65 (1), 361–392.
Clemen, R.T., 1989. Combining forecasts: a review and annotated bibliography. Int. J. Forecast. 5 (4), 559–583.
Collins, D.W., Kothari, S.P., Shanken, J., Sloan, R.G., 1994. Lack of timeliness and noise as explanations for the low contemporaneous return-earnings association. J. Account. Econ. 18 (3), 289–324.
Coqueret, G., 2022. Persistence in factor-based supervised learning models. J. Finance Data Sci. 8, 12–34.
De Moor, L., Sercu, P., 2013. The smallest firm effect: an international study. J. Int. Money Finance 32, 129–155.
Dechow, P.M., Sloan, R.G., Sweeney, A.P., 1995. Detecting earnings management. Account. Rev. 193–225.
Docherty, P., Hurst, G., 2018. Investor myopia and the momentum premium across international equity markets. J. Financ. Quant. Anal. 53 (6), 2465–2490.
Diebold, F., Mariano, R, 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13 (3), 253–263.
Dong, X., Li, Y., Rapach, D.E., Zhou, G., 2022. Anomalies and the expected market return. J. Finance 77 (1), 639–681.
Choi, D., Jiang, W., & Zhang, C. (2022). Alpha go everywhere: machine learning and international stock returns. Available at SSRN 3489679.
Drobetz, W., Hollstein, F., Otto, T., & Prokopczuk, M. (2021). Estimating security betas via machine learning. Available at SSRN 3933048.
Drobetz, W., Otto, T., 2021. Empirical asset pricing via machine learning: evidence from the european stock market. J. Asset Manag. 22 (7), 507–538.
Durnev, A., Morck, R., Yeung, B., Zarowin, P., 2003. Does greater firm-specific return variation mean more or less informed stock pricing? J. Account. Res. 41 (5), 797–836.
Ehsani, S., Linnainmaa, J.T., 2022. Factor momentum and the momentum factor. J. Finance 77 (3), 1877–1919.
Fama, E.F., French, K.R., 1993. Common risk factors in the returns on stocks and bonds. J. Financ. Econ. 33 (1), 3–56.
Fama, E.F., French, K.R., 2008. Dissecting anomalies. J. Finance 63 (4), 1653–1678.
Fama, E.F., French, K.R., 2012. Size, value, and momentum in international stock returns. J. Financ. Econ. 105 (3), 457–472.
Fama, E.F., French, K.R., 2017. International tests of a five-factor asset pricing model. J. Financ. Econ. 123 (3), 441–463.
Fama, E.F., French, K.R., 2018. Choosing factors. J. Financ. Econ. 128 (2), 234–252.
Fama, E.F., MacBeth, J.D., 1973. Risk, return and equilibrium: empirical tests. J. Polit. Econ. 81 (3), 607–636.
Feng, G., He, J., & Polson, N.G. (2018). Deep learning for predicting asset returns. arXiv preprint arXiv:1804.09314.
Feng, G., Polson, N., Xu, J., 2023. Deep learning in characteristics-sorted factor models. J. Financ. Quant. Anal. in press.
Filippou, I., Rapach, D., Taylor, M.P., & Zhou, G. (2020). Exchange rate prediction with machine learning and a smart carry portfolio. Available at SSRN 3455713.
Freyberger, J., Neuhierl, A., Weber, M., 2020. Dissecting characteristics nonparametrically. Rev. Financial Studies 33 (5), 2326–2377.
Gao, P., Parsons, C.A., Shen, J., 2018. Global relation between financial distress and equity returns. Rev. Financial Studies 31 (1), 239–277.
Goyal, A., Wahal, S., 2015. Is momentum an echo? J. Financ. Quant. Anal. 50 (6), 1237–1267.
Green, J., Hand, J., Zhang, F., 2017. The characteristics that provide independent information about average US monthly stock returns. Rev. Financial Studies 30, 4389–4436.
Griffin, J.M., Kelly, P.J., Nardari, F., 2010. Do market efficiency measures yield correct inferences? A comparison of developed and emerging markets. Rev. Financial Studies 23 (8), 3225–3277.
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. Rev. Financial Studies 33 (5), 2223–2273.
Gu, S., Kelly, B., Xiu, D., 2021. Autoencoder asset pricing models. J. Econom. 222 (1), 429–450.
Haddad, V., Kozak, S., Santosh, S., 2020. Factor timing. Rev. Financial Studies 33 (5), 1980–2018.
Han, Y., He, A., Rapach, D., & Zhou, G. (2023). Cross-sectional expected returns: new Fama-MacBeth regressions in the era of machine learning. Available at SSRN: https://ssrn.com/abstract=3185335 or doi:10.2139/ssrn.3185335.
Hanauer, M.X., Kalsbach, T., 2022. Machine learning and the cross-section of emerging market stock returns. Emerg. Markets Rev. in press.
Harvey, C.R., Liu, Y., Zhu, H., 2016. … and the cross-section of expected returns. Rev. Financial Studies 29 (1), 5–68.
Heaton, J.B., Polson, N.G., Witte, J.H., 2017. Deep learning for finance: deep portfolios. Appl. Stoch. Models Bus. Ind. 33 (1), 3–12.
Hollstein, F., 2022. The world of anomalies: smaller than we think? J. Int. Money Finance 129, 102741.
Hollstein, F., & Sejdiu, V. (2020). Probability distortions, collectivism, and international stock prices. Available at SSRN: https://ssrn.com/abstract=3737342 or doi:10.2139/ssrn.3737342.
Hong, H., Lim, T., Stein, J.C., 2000. Bad news travels slowly: size, analyst coverage, and the profitability of momentum strategies. J. Finance 55 (1), 265–295.
Hou, K., Xue, C., Zhang, L., 2015. Digesting anomalies: an investment approach. Rev. Financial Studies 28 (3), 650–705.
Hou, K., Xue, C., Zhang, L., 2020. Replicating anomalies. Rev. Financial Studies 33 (5), 2019–2133.
Jacobs, H., 2015. What explains the dynamics of 100 anomalies? J. Bank Financ. 57, 65–85.
Jacobs, H., 2016. Market maturity and mispricing. J. Financ. Econ. 122 (2), 270–287.
Jacobs, H., Müller, S., 2020. Anomalies across the globe: once public, no longer existent? J. Financ. Econ. 135 (1), 213–230.
Jensen, T.I., Kelly, B.T., Pedersen, L.H., 2022. Is there a replication crisis in finance? J. Finance in press.
Jiang, G., Lee, C., Zhang, Y., 2005. Information uncertainty and expected returns. Rev. Account. Stud. 10, 185–221.
Kelly, B.T., Malamud, S., Zhou, K., 2023. The virtue of complexity in return prediction. J. Finance in press.
Kelly, B.T., Pruitt, S., Su, Y., 2019. Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. 134 (3), 501–524.
Kim, S., Korajczyk, R.A., Neuhierl, A., 2021. Arbitrage portfolios. Rev. Financial Studies 34 (6), 2813–2856.
Koijen, R.S., Moskowitz, T.J., Pedersen, L.H., Vrugt, E.B., 2018. Carry. J. Financ. Econ. 127 (2), 197–225.
Kozak, S., Nagel, S., Santosh, S., 2020. Shrinking the cross-section. J. Financ. Econ. 135 (2), 271–292.
Lam, F.E.C., Wei, K.J., 2011. Limits-to-arbitrage, investment frictions, and the asset growth anomaly. J. Financ. Econ. 102 (1), 127–149.
Leippold, M., Wang, Q., Zhou, W., 2022. Machine learning in the Chinese stock market. J. Financ. Econ. 145 (2A), 64–82.
Leitch, G., Tanner, J.E., 1991. Economic forecast evaluation: profits versus the conventional error measures. Am. Econ. Rev. 580–590.
Lettau, M., Pelger, M., 2020a. Estimating latent asset-pricing factors. J. Econom. 218 (1), 1–31.
Lettau, M., Pelger, M., 2020b. Factors that fit the time series and cross-section of stock returns. Rev. Financial Studies 33 (5), 2274–2325.
Leuz, C., Nanda, D., Wysocki, P.D., 2003. Earnings management and investor protection: an international comparison. J. Financ. Econ. 69 (3), 505–527.
Linnainmaa, J.T., Roberts, M.R., 2018. The history of the cross-section of stock returns. Rev. Financial Studies 31 (7), 2606–2649.
Fama, E.F., 1970. Efficient capital markets: a review of theory and empirical work. J. Finance 25 (2), 383–417.
Liu, Q., Tao, Z., Tse, Y., Wang, C., 2022. Stock market prediction with deep learning: the case of China. Finance Res. Lett. 46, 102209.
McLean, R.D., 2010. Idiosyncratic risk, long-term reversal, and momentum. J. Financ. Quantit. Anal. 45 (4), 883–906.
McLean, R.D., Pontiff, J., 2016. Does academic research destroy stock return predictability? J. Finance 71 (1), 5–32.
Morck, R., Yeung, B., Yu, W., 2000. The information content of stock markets: why do emerging markets have synchronous stock price movements? J. Financ. Econ. 58 (1–2), 215–260.
Müller, K., & Schmickler, S. 2020. Interacting anomalies. Available at SSRN: https://ssrn.com/abstract=3646417 or doi:10.2139/ssrn.3646417.
Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55 (3), 703–708.
Novy-Marx, R., 2012. Is momentum really momentum? J. Financ. Econ. 103 (3), 429–453.
Novy-Marx, R., 2013. The other side of value: the gross profitability premium. J. Financ. Econ. 108 (1), 1–28.
O'Doherty, M., Savin, N.E., Tiwari, A., 2012. Modeling the cross section of stock returns: a model pooling approach. J. Financ. Quant. Anal. 47 (6), 1331–1360.

Rad, H., Low, R.K.Y., Miffre, J., & Faff, R.W. (2021). The commodity risk premium and neural networks. Available at SSRN 3816170.
Rapach, D.E., Zhou, G., 2020. Time-series and cross-sectional stock return forecasting: new machine learning methods. Machine Learning for Asset Management: New Developments and Financial Applications, pp. 1–33.
Rapach, D.E., Strauss, J.K., Tu, J., Zhou, G., 2019. Industry return predictability: a machine learning approach. J. Financ. Data Sci. 1 (3), 9–28.
Rapach, D.E., Strauss, J.K., Zhou, G., 2010. Out-of-sample equity premium prediction: combination forecasts and links to the real economy. Rev. Financial Studies 23 (2), 821–862.
Rapach, D.E., Strauss, J.K., Zhou, G., 2013. International stock return predictability: what is the role of the United States? J. Finance 68, 1633–1662.
Rasekhschaffe, K.C., Jones, R.C., 2019. Machine learning for stock selection. Financial Anal. J. 75 (3), 70–88.
Sadka, R., Scherbina, A., 2007. Analyst disagreement, mispricing, and liquidity. J. Finance 62 (5), 2367–2403.
Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. J. Finance 52 (1), 35–55.
Sloan, R.G., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? Account. Rev. 289–315.
Stambaugh, R.F., Yu, J., Yuan, Y., 2012. The short of it: investor sentiment and anomalies. J. Financ. Econ. 104 (2), 288–302.
Struck, C., Cheng, E., 2020. The cross section of commodity returns: a nonparametric approach. J. Financ. Data Sci. 2 (3), 86–103.
Timmermann, A., 2006. Forecast combinations. Handbook Econ. Forecast. 1, 135–196.
Titman, S., Wei, K.J., Xie, F., 2013. Market development and the asset growth effect: international evidence. J. Financ. Quant. Anal. 48 (5), 1405–1432.
Tobek, O., Hronec, M., 2021. Does it pay to follow anomalies research? Machine learning approach with international evidence. J. Financ. Markets 56, 100588.
Umutlu, M., Akdeniz, L., Altay-Salih, A., 2010. The degree of financial liberalization and aggregated stock-return volatility in emerging markets. J. Bank Financ. 34 (3), 509–521.
Watanabe, A., Xu, Y., Yao, T., Yu, T., 2013. The asset growth effect: insights from international equity markets. J. Financ. Econ. 108 (2), 529–563.
Zaffaroni, P., & Zhou, G. (2022). Asset pricing: cross-section predictability. Available at SSRN 4111428.
Zhang, X.F., 2006. Information uncertainty and stock returns. J. Finance 61 (1), 105–137.