Six-Factor Asset Pricing versus LSTM Forecasting in the Pakistan Stock Exchange

Malaika Nisar [1] **Muhammad Naveed Jan** [2*] **Usman Ayub** [3]

[1] Research Scholar, Department of Management Sciences, COMSATS University Islamabad, Abbottabad Campus, Khyber Pakhtunkhwa, Pakistan. Email: malaikanisar4@gmail.com

[2] Assistant Professor, Department of Management Sciences, COMSATS University Islamabad, Abbottabad Campus, Khyber Pakhtunkhwa, Pakistan. Email: naveedjan@cuiatd.edu.pk

[3] Professor, Department of Management Sciences, COMSATS University Islamabad, Abbottabad Campus, Khyber Pakhtunkhwa, Pakistan. Email: usmanayub@cuiatd.edu.pk

ABSTRACT: This research examines whether the six-factor asset pricing model or Long Short-Term Memory (LSTM) recurrent neural network is better suited to forecasting the portfolio excess return in the Pakistan Stock Exchange (PSX), which is a frontier market with relatively less history. The data used is monthly for 300 PSX companies classified into 30 factor sorted portfolios (Jan 1999 – Dec 2025, 312 months), where a six-factor time-series regression with Newey-West HAC errors is estimated, followed by joint pricing error significance test using Gibbons-Ross-Shanken (GRS) statistic. A portfolio-specific LSTM is trained on the same six factors using a fixed chronological split (196 training, 42 validations, 42 test months) and compared with the regression on an identical 42-month out-of-sample window using RMSE, MAE, out-of-sample R^2, hit rate, and three formal tests. The six-factor model explains substantial return variation (mean in-sample R^2 = .758), but the GRS test rejects the null of jointly zero pricing errors ($F = 12.251$, $p < .0001$; mean $|\alpha| = 0.749\%$). The LSTM shows no systematic incremental forecasting value: the regression achieves lower RMSE on 28 of 30 portfolios and a higher out-of-sample R^2 (.590 vs. −.022), with all three formal tests rejecting equal accuracy at $p < .0001$. The LSTM's limited value concentrates in momentum-sorted portfolios, consistent with permutation importance ranking momentum as the dominant learnable input. We interpret this as evidence that algorithmic flexibility does not automatically yield forecasting gains under limited training data and noisy returns, not as evidence against machine learning generally.

KEYWORDS: Asset Pricing, Six-factor Model, LSTM, Frontier Markets, Pakistan, Return Forecasting

Pages: 180 – 193
Volume: 5
Issue: 8 (August 2026)

Corresponding Author
Muhammad Naveed Jan
✉ naveedjan@cuiatd.edu.pk

Cite this Article: Nisar, M., Jan, M. N., & Ayub, U. (2026). Six-Factor Asset Pricing versus LSTM Forecasting in the Pakistan Stock Exchange. *The Regional Tribune, 5*(8), 180-193. https://doi.org/10.55737/trt/v-viii.406

Introduction

Forecasting equity returns and explaining their systematic determinants remain two of the central, and still only partially resolved, problems in empirical finance. Multifactor asset-pricing models — from the Capital Asset Pricing Model (Sharpe, 1964; Lintner, 1965) through the Fama and French (1993) three-factor model to its five- and six-factor extensions (Fama & French, 2015, 2018) — provide a theoretically grounded account of the systematic risk exposures that should determine expected returns. At the same time, a fast-growing literature applies machine-

learning methods, and recurrent neural networks in particular, to the same forecasting problem, motivated by their capacity to represent nonlinear and time-varying relationships that a static linear model cannot (Gu et al., 2020; Chen et al., 2024).

Most of the evidence informing the presumption that nonlinear machine-learning models improve on linear factor models comes from large, liquid, data-rich markets — principally the United States — where decades of complete firm-level history are available to train high-capacity models. Frontier markets present a materially different setting: shorter reliable data histories, thinner trading, higher idiosyncratic noise, and fewer firms with the continuous records needed to construct systematic risk factors. Whether the same presumption of machine-learning superiority holds once these constraints bind is an open empirical question, and one with direct relevance for how asset-pricing and forecasting research should be conducted outside the largest markets.

PSX can serve as an interesting platform for studying the above problems. First, the exchange is recognized as a frontier market; secondly, it has a history of continuous listings lasting many years yet relatively scarce; thirdly, there is an increasing yet small amount of multifactor asset pricing studies, mostly focused on the three- and five-factor models (Roy & Shijin, 2018, provide one of the first six-factor models outside Pakistan; five-factor evidence for PSX is available, whereas the six-factor model evidence coupled with benchmarked nonlinear forecasting is more scarce).

This research aims to achieve three related objectives. The first objective is to test the ability of the six-factor model – market, size, value, momentum, profitability, and investment – to explain the returns on 30 factor-sorted portfolios on the PSX for 312 months (January 1999–December 2025) through both the standard F-test and the joint test of the significance of pricing errors based on the GRS statistic. The second objective is to train an LSTM recurrent neural network using the six aforementioned factors to investigate whether a nonlinear sequence-aware model can capture return dynamics that are not explained by the linear model. The third and main objective of the paper is to compare both models using the same out-of-sample window, in a completely controlled manner, through various statistical tests.

The results may be easily stated. The six-factor model accounts for a great deal of the systematic variation in the returns on PSX portfolios, as is typical of international asset pricing data; however, a formal GRS test shows that the null of joint zero pricing errors for the 30 portfolios is rejected. The LSTM, which utilizes the same six factors and is tested in an equivalent 42-month out-of-sample period, does not deliver any systematic value relative to the linear baseline – it loses to the linear predictor in 28 of 30 portfolios, and three formal tests (paired t-test, Wilcoxon signed-rank test, and Diebold-Mariano-type test) all reject equality of the forecasts in favor of the linear model. The only area in which the LSTM does add value is momentum – the permutation feature importance measure reveals momentum as the by-far most important input into the LSTM, while the few portfolios where the LSTM does provide incremental value are in momentum portfolios.

We interpret this pattern as evidence consistent with a specific, scope-limited claim: under the training-sample size, factor information set, and evaluation design examined here, algorithmic flexibility does not automatically translate into superior forecasting accuracy when the underlying return process is noisy, and the reliably usable estimation history is short — a condition more likely to bind in frontier markets than in the large, long-history markets where most machine-learning-in-asset-pricing evidence originates. We do not interpret this as evidence that machine learning is ineffective in general, nor that linear factor models are universally superior; both claims would overreach what a single market, single architecture, and single evaluation window can support.

Literature Review and Theoretical Foundation

Multifactor Asset Pricing

The CAPM (Sharpe, 1964; Lintner, 1965) posited that market risk is the only systematic driver of expected return. Systematic deviations from this model, in particular, the size effect (Banz, 1981) and the value effect (Fama & French, 1992), led Fama & French (1993) to their three-factor specification which adds the size (SMB) and value (HML) factors, and Carhart's (1997) inclusion of the momentum (WML) factor, based on Jegadeesh & Titman (1993) findings that past winners perform better than past losers for periods between 3 and 12 months. Fama & French (2015) added the profitability (RMW) and the investment (CMA) factors, inspired by the logic of the dividend discount model that, ceteris paribus, higher expected investment implies lower expected return, and formalized independently in the q-theory framework by Hou et al. (2015). In Fama & French (2018), the authors put forward a spanning-test approach to choosing among different factor specifications, and in this framework, six-factor formulations involving the size, value, profitability, investment, and momentum factors alongside the market factor (see e.g. Roy & Shijin, 2018) appear to be an empirical extension, not a canonical choice.

Factor Models in Emerging and Frontier Markets

However, the transportability of factor premia from the developed world is not a given process. While the value and profitability factors, in particular, have delivered relatively inconclusive findings outside the developed markets – in some cases even significantly lower or with an opposite sign – size and momentum effects have been found to be quite consistent across other institutional settings. Such asymmetry is economically relevant because it implies that the theories of the value and profitability factors, which involve mechanisms such as limits to arbitrage on distress risk, speed of mispricing adjustment, analyst coverage, etc., rely heavily on certain characteristics of the institutions of the developed world (such as depth of the markets, short selling restrictions, etc.) which are absent in frontier countries. Such considerations motivate the approach of considering factor premia magnitude, not just its sign, as a country-specific question, not a matter of direct transportability of parameters – the position taken by this paper.

Pakistan Stock Exchange Evidence

Research related to asset pricing on PSX has evolved from CAPM and three factors towards five- and six-factor models. One of the consistent findings in such studies, confirmed by the current study's own descriptive results, is that market effect and size effect tend to be significant and large, whereas the value effect tends to be insignificant and close to zero, contrary to the value premium found in studies of developed markets. Research in the area of six factors along with machine learning forecasts for PSX has remained relatively limited; this is the particular gap filled by this study.

Machine Learning in Empirical Asset Pricing

There is a significant body of literature using machine learning techniques – including regularized linear regressions, tree ensembles, and neural networks – for the purposes of return prediction in cross-sectional and time series environments. Gu et al. (2020) offer a well-known comparative assessment of numerous machine learning approaches within the cross-section of the U.S., finding, in general, out-of-sample improvement over linear benchmarks across many of the techniques, especially tree ensembles and neural networks trained on sufficiently sized datasets. This approach has been followed by Chen et al. (2024), who develop and evaluate a deep-learning constrained asset pricing model under no-arbitrage assumptions on a long history of data on the U.S. market. What all these works have in common and what should be clearly stated is the fact that they use an extensive evidence base

collected in situations where there are no severe restrictions on the sample size for model training. The question of whether the performance of models would improve in other cases is what this paper attempts to answer.

LSTM/RNN Forecasting

Long Short-Term Memory Network (LSTM) (Hochreiter & Schmidhuber, 1997) overcomes the vanishing gradient problem that was a major drawback in earlier neural networks and thus is particularly appropriate for financial data due to its sequence-based nature. Fischer and Krauss (2018) provided an influential early application to daily equity return prediction, reporting economically meaningful outperformance over standard benchmarks on a large U.S. universe with an extensive training history. More directly comparable to the present study, Afzal et al. (2025) apply an LSTM to a Fama-French six-factor input specification on the Shenzhen (China) A-share market and — notably, given that market's status as a large and liquid emerging rather than frontier market — likewise report that model complexity alone does not guarantee forecasting improvement, using a random-forest surrogate to characterize the nonlinear structure the LSTM captures. That a broadly similar qualification emerges even in a considerably larger and more liquid market than the PSX strengthens, rather than weakens, the plausibility of the present study's frontier-market findings: if data-availability constraints matter even in Shenzhen, they should matter at least as much, and plausibly more, in a market with a shorter reliable history and thinner trading.

Econometric versus Machine-Learning Forecasting Comparisons

Direct, fully controlled comparisons between linear factor models and machine-learning forecasts — using identical portfolios, an identical evaluation window, and multiple formal statistical tests rather than a single accuracy metric — remain less common than either literature considered in isolation. Much of the machine-learning-in-asset-pricing literature reports a machine-learning model's own performance metrics without an equivalently rigorously specified and estimated linear counterfactual evaluated under matched conditions; conversely, much of the classical asset-pricing literature does not engage with machine-learning forecasting comparisons at all. This study's design — identical portfolios, identical 42-month test window, and three independent formal comparison tests (paired t, Wilcoxon signed-rank, Diebold–Mariano-style) — is intended to occupy this specific, comparatively under-populated methodological space.

Research Gap and Positioning

Three gaps motivate this study. First, six-factor (profitability- and investment-augmented) asset-pricing evidence specific to the PSX is thinner than the existing three- and five-factor literature. Second, studies combining Fama-French-style factor models with LSTM forecasting exist for some emerging markets — including the Shenzhen evidence discussed above — but are sparse for frontier markets, and rarely implement a fully controlled, identical-portfolio, identical-window comparison with multiple formal significance tests. Third, the broader machine-learning-in-asset-pricing literature (Gu et al., 2020; Chen et al., 2024) is built largely on data-rich developed-market panels; systematic evidence on how frontier-market training-sample constraints affect the realized, as distinct from theoretical, forecasting advantage of nonlinear models over linear benchmarks is limited.

Relative to the closest comparators identified in this review, this study differs specifically as follows: relative to Gu et al. (2020) and Chen et al. (2024), it examines a frontier rather than developed market and a substantially shorter reliable estimation history; relative to Afzal et al. (2025), it examines a frontier rather than large-emerging market, a six-factor rather than an unspecified factor-count input, and adds a formal GRS test of the linear benchmark's own pricing adequacy, which that study does not report. To the best of our knowledge, no existing

published study combines a six-factor PSX asset-pricing specification with a formally benchmarked, GRS-validated LSTM comparison using an identical out-of-sample evaluation protocol for both models; we make this claim narrowly, about this specific combination of market, factor count, and comparison design, rather than as a broader priority claim.

Data and Methodology

Sample and Data

The sample includes 300 PSX-listed companies having uninterrupted monthly pricing and accounting information for the period January 1999 through December 2025 (312 months), which is based on official PSX price data and the companies' financial statements, with State Bank of Pakistan's Treasury bill rate utilized as the risk-free rate estimate. The selection of the companies is made through a purposive (non-random sampling) method due to continuity in the data across the entire sample period, with consideration of a variety of sectors so as not to concentrate in one sector.

This criterion for selection becomes important as the requirement of 26 years of continuous and complete data may lead to the selection of only those firms that survived through the period and have good record-keeping practices, which may be associated with size and financial stability. We do not have a basis to quantify or correct for this in the present data, and we do not claim that it demonstrably biases the reported results; we flag it as a limitation of the sampling design.

Portfolio Construction

Following Fama and French (1993, 2015), stocks are sequentially sorted on market capitalization (Small/Big) and, within each size group, on book-to-market ratio (High/Low), cumulative 6–12-month formation-period returns (Winner/Loser), operating profitability (Robust/Weak), and asset growth (Conservative/Aggressive), yielding 30 factor-sorted portfolios (five sort families: Size × Value, Size × Momentum, Size × Profitability, Size × Investment, and Value × Momentum). Monthly portfolio excess returns (portfolio return minus the risk-free rate) constitute the dependent variable throughout.

Factor Construction

Six systematic risk factors are constructed: market excess return (MKT), size (SMB), value (HML), momentum (WML), profitability (RMW), and investment (CMA), following standard Fama-French/Carhart construction methods:

$MKT(t) = R_{m,t} - Rf_{,t}$ *(Eq. 1)*

SMB(t) = (1/3)[Avg(Small)] − (1/3)[Avg(Big)] *(Eq. 2)*

HML(t) = (1/2)[Avg(High B/M)] − (1/2)[Avg(Low B/M)] *(Eq. 3)*

$WML(t) = Avg(Winners)_t - Avg(Losers)_t$ *(Eq. 4)*

RMW(t) = (1/2)[Avg(Robust)] − (1/2)[Avg(Weak)] *(Eq. 5)*

CMA(t) = (1/2)[Avg(Conservative)] − (1/2)[Avg(Aggressive)] *(Eq. 6)*

A necessary data-limitation disclosure: complete firm-level operating-profitability and asset-growth data meeting SECP reporting standards were not available for the full 1999–2025 window for all sampled firms. RMW and CMA are therefore constructed as regression-estimated proxy factors rather than being built directly from the canonical Fama and French (2015) firm-level definitions. This is not a cosmetic distinction: proxy RMW and CMA should not be interpreted as numerically equivalent to the RMW/CMA reported in developed-market studies, and their

loadings, premia, and their machine-learning permutation importance should be read with this construction difference in mind.

Six-Factor Regression and the GRS Test

For each portfolio, excess returns are estimated for the following six factors using OLS regression with Newey-West HAC standard errors (with six lags):

$$R_{p,t} - Rf_{,t} = \alpha_p + \beta 1 \cdot MKT_t + \beta 2 \cdot SMB_t + \beta 3 \cdot HML_t + \beta 4 \cdot WML_t + \beta 5 \cdot RMW_t + \beta 6 \cdot CMA_t + \varepsilon_{p,t} \quad (Eq.\ 7)$$

estimated using the full 312 months of data (January 1999 to December 2025). Then the GRS statistic from Gibbons et al. (1989) jointly tests H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_{30} = 0$, meaning whether the six-factor model prices the cross section of the 30 portfolios without any unexplained component.

Regarding the span of time period: the full 312-month period is used for this explanatory (in-sample) analysis. The forecast comparison with the LSTM involves the common 42-month period out-of-sample analysis mentioned, including a separate computation of out-of-sample FF6 R^2 in the same period, so as not to confuse the two analyses.

With regard to the span of loadings reported, the sign, size, and significance of the market beta and the momentum factor loading in all 30 portfolios, as reported in their entirety in the underlying source material.

Regression Diagnostics

Four diagnostic tests assess classical OLS assumptions: the Durbin–Watson statistic and a Ljung–Box Q(12) test for serial correlation, a Breusch–Pagan test for heteroskedasticity, and a Jarque–Bera test for residual normality, reported for all 30 portfolios.

LSTM Architecture and Evaluation Design

A portfolio-specific two-layer LSTM (64 units → dropout 0.2 → 32 units → dropout 0.2 → 16-unit dense layer → single output) is trained independently for each of the 30 portfolios, using the same six factors reorganized into overlapping 12-month rolling input sequences (each sample uses months t−11 through t to predict the excess return at t+1). Inputs are min-max normalized using training-set parameters only, applied subsequently to the validation and test sets to avoid look-ahead leakage.

Data are split by a single, fixed chronological partition — 196 training months (January 1999–April 2015), 42 validation months (May 2015–October 2018), and 42 test months (November 2018–April 2022) — trained with the Adam optimizer (learning rate 0.001), mean squared error loss, batch size 16, and early stopping (patience 15 epochs on validation loss, restoring best weights; mean convergence after approximately 19 epochs across the 30 networks).

We describe this explicitly as a fixed chronological (out-of-time) split rather than a rolling-window re-estimation scheme: the reported design is a single train/validation/test partition, not repeated re-estimation across successive or expanding forecast origins. This distinction matters for interpreting the LSTM's out-of-sample performance and is taken up again as a limitation and future-research direction.

Forecast Evaluation Metrics

Both models are evaluated on the identical 42-month test window (November 2018–April 2022) using RMSE, MAE, out-of-sample R^2 (relative to the historical-mean benchmark; Campbell & Thompson, 2008), and directional

hit rate. MAPE is also reported by the underlying source data but is not used as a primary criterion here: because monthly portfolio excess returns are frequently close to zero and change sign, percentage-based error measures become numerically unstable and can produce very large values that do not indicate a comparably large economic forecasting failure. We note this once and do not feature MAPE in the headline comparison tables.

Model Comparison Tests

Three complementary tests assess whether observed performance differences are statistically distinguishable from sampling variation across the 30 paired portfolio-level observations: a paired t-test (parametric, assumes approximate normality), a Wilcoxon signed-rank test (nonparametric, distribution-free), and a squared-error-loss differential t-test adapted from the Diebold and Mariano (1995) forecast-comparison framework.

Cross-sectional dependence in the comparison tests. The 30 portfolio-level paired observations used in the formal comparison tests are not fully mutually independent: portfolios constructed from overlapping sort dimensions (e.g., all portfolios sorted into the "Small" size group) share common exposure to at least one sorting variable. The paired t-test and Diebold–Mariano-style test both nominally assume independent observations; the reported test statistics should therefore be read as informative, but not as having exactly the nominal degrees of freedom implied by treating the 30 portfolios as independent draws. The Wilcoxon signed-rank test, being distribution-free with respect to the underlying error distribution, is less sensitive to this concern but does not itself correct for cross-sectional dependence in the observations. We report all three tests together specifically so that the reader can assess whether the conclusion (regression outperforms LSTM) is sensitive to this issue: because all three tests agree in direction and each independently rejects equal accuracy at $p < .0001$, the qualitative conclusion is unlikely to be an artifact of the dependence structure alone, though the exact p-values should be interpreted with this caveat in mind.

Empirical Results

Descriptive Statistics

Table 1

Descriptive Statistics of PSX Six-Factor Returns (Jan 1999–Dec 2025, T = 312 monthly observations, % returns)

Factor	Mean (%)	SD (%)	Min (%)	Max (%)	Ann. Sharpe
MKT	0.8053**	6.3110	−20.32	26.09	0.442
SMB	0.4218**	3.4419	−10.28	10.09	0.425
HML	0.2302	3.2795	−9.63	9.40	0.243
WML	0.5967*	5.5250	−13.27	16.61	0.374
RMW†	0.4538**	2.7911	−7.26	9.32	0.563
CMA†	0.2438*	2.4636	−6.64	10.68	0.343

Note. */**/*** denote significance at the 10/5/1% levels (Newey–West HAC t-test, H_0: mean = 0). † RMW and CMA are regression-estimated proxy factors; their comparability with canonical Fama-French RMW/CMA should not be assumed.

The market factor carries the largest average premium and by far the highest volatility, consistent with its expected role as the dominant systematic risk source in an equity market. Size and momentum premia are positive and statistically significant; the value premium (HML), by contrast, is small and statistically indistinguishable from zero — a pattern consistent with prior PSX evidence that the classical book-to-market effect documented in developed markets is comparatively weak in this market. The pairwise correlations among the six factors are

uniformly small (maximum |r| = .103, between SMB and WML), indicating limited multicollinearity risk for the regression specification and near-orthogonal information content across the six sequences supplied to the LSTM.

Factor Correlations

The correlation structure summarized above supports the regression specification without further remediation; full pairwise correlations are available in supplementary material.

Six-Factor Regression Results

The values of in-sample R^2 for the 30 portfolios vary between .664 and .873 (mean = .758), and the market beta is positive, close to one, and highly significant in all portfolios (values range from .807 to 1.093; p<.001), thereby providing evidence on the dominance of the market factor's role. There is an internally consistent pattern in the cross-section with regard to the momentum loading – portfolios sorted according to momentum have winners with positive loading and losers with negative loading; moreover, several absolute loading values exceed 0.6, and their t-statistics exceed 12 in absolute value, which proves the construction and the pricing of the momentum factor. 25 of the 30 estimated pricing errors (alphas) are statistically significant, negative, and equal to between -0.30% and -1.51% per month (at the 1% significance level); only five portfolios (Size × Value-Big/High, both Size × Momentum "Winner" portfolios divided by size, Size × Profitability-Small/Robust, and Value×Momentum-High/Winner) have alphas statistically insignificant from zero.

GRS Test Results

Table 2

GRS Test of Joint Alpha Significance (Gibbons, Ross, & Shanken, 1989)

Sort Family	N	GRS F-stat	p-value	Mean |α| (%)	Decision
Size × Value	6	12.058	< .0001	0.701	Rejected
Size × Momentum	6	8.408	< .0001	0.613	Rejected
Size × Profitability	6	13.920	< .0001	0.738	Rejected
Size × Investment	6	15.285	< .0001	0.746	Rejected
Value × Momentum	6	15.257	< .0001	0.946	Rejected
All 30 (Joint)	30	12.251	< .0001	0.749	Rejected

The joint GRS test on the entire collection of 30 portfolios rejects the null hypothesis of joint zero pricing errors (F = 12.251, p < .0001). This is clear and consistent with our language throughout this paper – we reject, not accept, the null hypothesis. While the six-factor model captures a substantial amount of systematic variation in returns, the cross-section of PSX portfolio returns is not completely priced by the model, and there is a significant unexplained part to be accounted for, especially in the case of the Value × Momentum family of portfolios (highest F-statistic and highest mean absolute α).

A meaningful interpretation, rather than a purely statistical one, should be given to the aggregation of the largest F-statistics of the GRS test and the largest mean absolute |α| in the Value × Momentum category. Indeed, Value × Momentum portfolio sorts are based on two factors whose relationship is uncertain theoretically. It is known from a broader perspective that momentum portfolios are vulnerable to rapid reversals after periods of market stress, whereas value strategies exhibit the tendency to experience lengthy valuation. In turn, a six-factor model with fixed (non-time-varying) coefficients cannot reflect any of the above dynamics if they operate on a different frequency

than the monthly one. The above statement is in line with, even though not proved by, about two portfolios that outperform the regression by LSTM, being precisely the Value × Momentum ones on the book-to-market extremes.

Diagnostic Results

Table 3

Summary of Regression Diagnostic Tests (All 30 Portfolios)

Test	Benchmark	Result	Interpretation
Durbin–Watson	≈ 2.0 = none	Mean 2.004; range 1.785–2.301	30/30 within acceptable range
Ljung–Box Q(12)	p > .05 = no correlation	28/30 portfolios p > .05	Residuals largely free of autocorrelation
Breusch–Pagan	p > .05 = homoscedastic	29/30 portfolios p > .05	Newey–West HAC appropriately applied
Jarque–Bera	p > .05 = normal	29/30 portfolios p > .05	Residuals approximately normal

Classical OLS assumptions are satisfied in the large majority of the 30 portfolio-level models, supporting the validity of the reported factor loadings, pricing errors, and the GRS joint test.

LSTM Forecasting Results

Across the 30 portfolio-specific networks, evaluated on the 42-month common test window (November 2018–April 2022): mean RMSE = .0729, mean MAE = .0569, mean out-of-sample R^2 = –.0223, and mean directional hit rate = 44.0%. Ten of 30 portfolios (33%) achieve positive out-of-sample R^2; the remaining 20 are negative. The mean hit rate (44.0%) is below the 50% level expected from random directional guessing, indicating the network's directional predictions are, on average across these portfolios, not more informative than chance — and marginally worse. MAPE values are large and in some cases exceed 500% for individual portfolios, we attribute this to near-zero-denominator instability rather than to a comparably severe economic forecasting failure, and do not treat it as a primary metric.

Feature Importance

Table 4

Mean Permutation Feature Importance (RMSE), Averaged Across 30 Portfolio LSTM Models

Factor	Mean ΔRMSE
MKT	0.0011
SMB	–0.0029
HML	0.0010
WML	0.0093
RMW	–0.0045
CMA	0.0061

Momentum (WML) is by a wide margin the most influential input, with mean importance roughly 1.5 times that of the next-ranked factor (CMA). SMB and RMW show negative mean ΔRMSE, meaning that randomly shuffling their input sequences did not, on average, degrade — and in some cases modestly improved — the network's forecasting performance. We interpret this cautiously: negative permutation importance in this context most plausibly reflects limited or unstable incremental predictive contribution from these variables under the specific LSTM configuration examined (finite-sample effects, feature redundancy, or interaction effects with other inputs),

not a demonstrated negative economic risk premium associated with either factor. This interpretation should also be read alongside the RMW proxy-construction caveat.

Direct Model Comparison

Table 5

Aggregate Performance Comparison — Six-Factor Regression vs. LSTM (All 30 Portfolios, Common 42-Month Test Window)

Metric	Six-Factor Regression	LSTM	Superior Model
Mean RMSE	0.04169	0.07292	Regression (lower better)
Mean MAE	0.03119	0.05695	Regression (lower better)
Mean MSE	0.002291	0.005459	Regression (lower better)
Mean OOS R^2	0.5898	−0.0223	Regression (higher better)
Mean hit rate (%)	82.1	44.0	Regression (higher better)
Portfolios OOS $R^2 > 0$	28/30	10/30	Regression
Portfolios won (RMSE)	28/30 (93.3%)	2/30 (6.7%)	Regression

Table 6

Formal Statistical Comparison Tests (N = 30 Paired Observations, Two-Sided)

Test	Metric	Statistic	p-value	Conclusion
Paired t-test	RMSE	−7.919	< .0001	Regression significantly lower RMSE
Wilcoxon signed-rank	RMSE	W = 7.0	< .0001	Regression significantly lower RMSE
Paired t-test	MAE	−8.728	< .0001	Regression significantly lower MAE
Wilcoxon signed-rank	MAE	W = 6.0	< .0001	Regression significantly lower MAE
Diebold–Mariano-style	MSE	−8.807	< .0001	Regression significantly lower MSE
Paired t-test	OOS R^2	10.253	< .0001	Regression significantly higher OOS R^2

All six tests reject equal predictive accuracy at $p < .0001$, and the nonparametric Wilcoxon results corroborate the parametric paired t-tests, indicating the conclusion is not an artifact of a normality assumption. Two exceptions to the dominant regression out-of-sample R^2 — Value × Momentum strategies where the regression R^2 is negative — are both characterized by extreme sorting on the book-to-market ratio and imply localized regime sensitivity that neither framework can capture, rather than a general superiority of LSTM.

The performance comparison provided herein is conducted based on a fixed 42-month out-of-sample testing period (November 2018 - April 2022). This testing period coincidentally includes the beginning and the early recovery stage of the market shock related to the emergence of the COVID-19 pandemic. It cannot be excluded from this experiment design itself that the extent of the regression's superior performance against LSTM may be confined to the particular testing period under consideration. The three formal tests confirm the regression's advantage is not a product of chance within this window; they cannot, by construction, confirm that the same advantage would be observed in a different window. This is a design limitation of the single fixed-split evaluation approach itself, not of the specific statistical tests applied to it, and is discussed further, along with the rolling-origin extension it motivates.

Discussion

This result supports three related but distinct conclusions instead of a global conclusion regarding “linear vs. nonlinear” models. First, the six-factor model explains much of PSX portfolio returns (in-sample mean $R^2 = .758$)

while leaving behind an unpriced element that is statistically significant (GRS test) - in line with the general conclusion from asset pricing literature that no linear factor model can price all the assets under consideration in its entirety. We do not read this as evidence that the six-factor specification is inadequate for describing PSX systematic risk; rather, it is the expected signature of a well-specified but necessarily incomplete linear model, and the magnitude of the remaining pricing error (mean $|\alpha|$ = 0.749% monthly) provides a concrete benchmark against which future extensions can be measured.

Second, the LSTM does not demonstrate systematic incremental forecasting value over the linear benchmark under the specific design examined here. Several non-exclusive explanations are plausible, and we deliberately present them as candidate interpretations rather than as demonstrated mechanisms, since disentangling them would require additional experiments beyond this study's scope. The 196-month training window used for each of the 30 independently trained networks is short relative to typical practice for training stable deep-learning models, and the single fixed evaluation window means we cannot separate a training-sample-size explanation from a test-window-specific explanation using this design alone. The portfolio excess-return series themselves are also noisy and close to mean-zero with limited apparent autocorrelation structure at the monthly frequency, a signal-to-noise environment that is difficult for any model, linear or nonlinear, to forecast well; the linear model's advantage may partly reflect that a simple, low-variance estimator is better suited to a low-signal environment than a higher-variance, higher-capacity one, independent of any specific failure of the LSTM architecture chosen here. Finally, because no hyperparameter search beyond the single reported architecture is documented, the result should be read as evidence about this LSTM specification under this training regime, not as a bound on what any LSTM configuration could achieve on this data.

Third, the LSTM's one area of genuine, if narrow, value-add — momentum-sorted portfolios, where permutation importance identifies momentum as the dominant input, and where the network's small number of "wins" over the regression concentrate — is consistent with a broader literature associating momentum returns with state-dependent, time-varying investor behavior (Jegadeesh & Titman, 1993). We describe the LSTM's momentum-related performance as consistent with this behavioral account rather than as evidence this study demonstrates causally; the regression-versus-LSTM comparison itself tests forecasting accuracy, not behavioral mechanisms. The concentration of both the GRS test's largest pricing error and the LSTM's clearest (if still exceptional) outperformance in Value×Momentum-adjacent portfolios is a pattern we flag as suggestive of shared underlying dynamics — momentum-related regime sensitivity — that neither the linear model's constant coefficients nor this specific LSTM's fixed 12-month lookback window fully resolve.

Taken together, the central interpretive theme we draw is that greater algorithmic flexibility does not necessarily produce greater forecasting accuracy when the information set is limited, the underlying return process is noisy, and the training history is short. We treat this as a study-specific interpretation, bounded by the one market, one factor input set, one LSTM architecture, and one evaluation window examined here — not a universal claim about the relative merits of econometric and machine-learning methods in asset pricing generally, and not a claim that a different architecture, longer history, or rolling-origin evaluation design would necessarily produce the same result.

Implications

For researchers: these results argue for benchmarking any machine-learning forecasting claim in a frontier-market setting against a well-specified linear factor model estimated and evaluated under identical conditions, rather than reporting machine-learning performance in isolation. The reusable two-stage pipeline documented here (factor/portfolio construction → linear benchmark with GRS/diagnostics → portfolio-specific LSTM → matched

out-of-sample comparison with multiple formal tests) is offered as a template for replication on other frontier markets or periods, independent of this study's specific empirical result.

For model selection in practice: for the monthly-horizon forecasting problem examined here, the six-factor specification is currently the more accurate forecasting tool for the large majority of PSX portfolios, particularly Size × Profitability and Size × Investment portfolios, where its out-of-sample R^2 frequently exceeds .90. We deliberately frame this as a model-selection and forecasting-accuracy conclusion, not an investment recommendation: no trading rule, transaction-cost analysis, turnover measure, or risk-adjusted investment-strategy evaluation was implemented in this study, and none of the results reported here should be read as evidence of exploitable trading profitability.

For machine-learning practitioners more broadly: the pattern of partial, momentum-concentrated LSTM success alongside overall underperformance is a useful diagnostic — the presence of some learnable nonlinear signal and the absence of net forecasting improvement are not mutually exclusive findings, and reporting only the aggregate comparison would obscure the more specific momentum-related result.

Limitations and Future Research

Several limitations bound the interpretation of these results. First, RMW and CMA are regression-estimated proxy factors, constructed in the absence of complete firm-level operating-profitability and asset-growth data for the full sample period; their loadings, premia, and permutation-importance results should be interpreted with this construction difference from canonical Fama-French RMW/CMA in mind. Second, the purposive, continuous-data sampling criterion used to construct the 300-firm sample may introduce a data-availability or survivorship-adjacent selection effect, favoring firms that survived and maintained clean records over the full 26-year window; we flag this possibility without claiming it is demonstrated or quantified in the present data.

Third, the LSTM evaluation uses a single, fixed chronological train/validation/test split rather than a rolling-origin or expanding-window re-estimation design; the reported out-of-sample results are therefore specific to this one evaluation window and this one partition, and we cannot separate a training-sample-size explanation for the LSTM's underperformance from a test-window-specific explanation using this design alone. Fourth, each of the 30 portfolio-specific LSTM networks was trained on a comparatively short 196-month window with a fixed architecture; no hyperparameter search beyond this specification is documented, and the results are specific to this architecture. Fifth, no transaction costs, turnover, or trading-strategy evaluation was conducted, so no claim of investment profitability follows from either model's forecasting accuracy.

Sixth, the sample period spans major macroeconomic and market shocks (the global financial crisis, the COVID-19 shock) whose effects on both models' relative performance are not separately decomposed here. Seventh, the 30 portfolio-level paired observations used in the formal comparison tests are not fully mutually independent, given shared sort-dimension exposure across portfolios; the reported p-values should be read as informative but not as reflecting fully independent observations.

The restrictions on this study suggest clear extensions for future research. Using an expanding window approach to testing the LSTM's performance relative to the chronological split here could determine if the network's results are dependent on the chronological split. Different models (GRU, Transformer, or an attention-based sequence model) could be benchmarked against each other using the same protocol for controlled comparison. Hybrid models that train a non-linear model on the residuals generated from a linear model (in this case, the GRS-verified residual mispricing) as opposed to on the raw returns are a possible extension based on this study's results: the problem in

the six-factor model that this study highlights (residual alphas) is exactly what a non-linear residual-forecasting model would aim to solve.

Conclusion

The present paper aimed at investigating which model is more explanatory of excess returns of a portfolio at the Pakistan Stock Exchange, that is, the six-factor asset pricing model or the LSTM recurrent neural network using the six systematic risks. The six-factor model explains a substantial share of return variation (mean in-sample R^2 = .758) but leaves a statistically significant unpriced component, formally confirmed by rejection of the GRS joint-alpha null hypothesis ($F = 12.251$, $p < .0001$).

The LSTM, trained on the same six factors and evaluated against the regression benchmark on an identical 42-month out-of-sample window, does not demonstrate systematic incremental forecasting value under the single architecture and single fixed-split evaluation design examined here: the regression outperforms it on 28 of 30 portfolios, with three independent formal tests confirming the difference is statistically significant, net of the cross-sectional-dependence caveat. LSTM's narrow area of genuine value — momentum-related portfolios — is consistent with momentum's documented behavioral and time-varying character.

We interpret these findings as evidence that, in a frontier-market setting with a comparatively short, reliable estimation history and noisy returns, algorithmic flexibility does not automatically translate into forecasting gains over a well-specified linear factor model — a scope-limited, evidence-specific conclusion about this study's design, not a general claim about the relative merits of econometric and machine-learning methods in asset pricing.

References

Afzal, T., Afridi, M. A., & Jan, M. N. (2025). Integrating LSTM with Fama-French six-factor model for predicting portfolio returns: Evidence from Shenzhen stock market, China. *Data Science in Finance and Economics*, *5*(2), 177-204. https://doi.org/10.3934/dsfe.2025009

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3-18. https://doi.org/10.1016/0304-405x(81)90018-0

Campbell, J. Y., & Thompson, S. B. (2007). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, *21*(4), 1509-1531. https://doi.org/10.1093/rfs/hhm055

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

Chen, L., Pelger, M., & Zhu, J. (2024). Deep learning in asset pricing. *Management Science*, *70*(2), 714-750. https://doi.org/10.1287/mnsc.2023.4695

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263. https://doi.org/10.1080/07350015.1995.10524599

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427. https://doi.org/10.2307/2329112

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3-56. https://doi.org/10.1016/0304-405x(93)90023-5

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

Fama, E. F., & French, K. R. (2018). Choosing factors. *Journal of Financial Economics*, *128*(2), 234-252. https://doi.org/10.1016/j.jfineco.2018.02.012

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669. https://doi.org/10.1016/j.ejor.2017.11.054

Gibbons, M. R., Ross, S. A., & Shanken, J. (1989). A test of the efficiency of a given portfolio. *Econometrica*, *57*(5), 1121. https://doi.org/10.2307/1913625

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hou, K., Xue, C., & Zhang, L. (2014). Digesting anomalies: An investment approach. *Review of Financial Studies*, *28*(3), 650-705. https://doi.org/10.1093/rfs/hhu068

Jegadeesh, N., & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65-91. https://doi.org/10.1111/j.1540-6261.1993.tb04702.x

Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, *47*(1), 13. https://doi.org/10.2307/1924119

Roy, R., & Shijin, S. (2018). A six-factor asset pricing model. *Borsa Istanbul Review*, *18*(3), 205-217. https://doi.org/10.1016/j.bir.2018.02.001

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425. https://doi.org/10.2307/2977928