
TEMPORAL KOLMOGOROV-ARNOLD NETWORKS (T-KAN) FOR HIGH-FREQUENCY LIMIT ORDER BOOK FORECASTING: EFFICIENCY, INTERPRETABILITY, AND ALPHA DECAY

Ahmad Makinde

Undergraduate Student, University of Bristol
Independent Researcher
Ahmad.makinde.2025@bristol.ac.uk

This research was conducted independently of the University of Bristol.

ABSTRACT

High-Frequency trading (HFT) environments are characterised by large volumes of limit order book (LOB) data, which is notoriously noisy and non-linear. Alpha decay represents a significant challenge, with traditional models such as DeepLOB losing predictive power as the time horizon (k) increases. In this paper, using data from the FI-2010 dataset, we introduce Temporal Kolmogorov-Arnold Networks (T-KAN) to replace the fixed, linear weights of standard LSTMs with learnable B-spline activation functions. This allows the model to learn the 'shape' of market signals as opposed to just their magnitude. This resulted in a 19.1% relative improvement in the F1-score at the $k = 100$ horizon. The efficacy of T-KAN networks cannot be understated, producing a **132.48%** return compared to the **-82.76%** DeepLOB drawdown under 1.0 bps transaction costs. In addition to this, the T-KAN model proves quite interpretable, with the 'dead-zones' being clearly visible in the splines. The T-KAN architecture is also uniquely optimized for low-latency **FPGA implementation** via High level Synthesis (HLS). The code for the experiments in this project can be found at <https://github.com/AhmadMak/Temporal-Kolmogorov-Arnold-Networks-T-KAN-for-High-Frequency-Limit-Order-Book-Forecasting>

Keywords Limit Order Book (LOB) · Alpha Decay · FI-2010 Dataset · Temporal Koolmogorov-Arnold Network (T-KAN) · Interpretability · FPGA

1 Introduction

The modeling and prediction of price dynamics in the Limit Order Book (LOB) are fundamental challenges in quantitative finance and market microstructure [1, 2]. Unlike low-frequency data, the LOB is a high-dimensional, discrete-event dynamic system where the latent state of supply and demand is shown via the placing and cancellation of orders across multiple price levels [3]. The LOB state at time t can be represented as a vector $\mathcal{L}_t = \{P_t^{(i)}, V_t^{(i)}\}_{i=-n'}^n$ where P and V represent the price and volume at level i , where positive and negative indices denote ask and bid sides, respectively.

In this state space, the **Auction Phase** is important. The phase is characterized by intense price discovery and structural liquidity shifts, where changing from a closed-call auction to continuous trading causes high-volatility regimes [4]. In order to forecast accurately, this regime requires the model to be able to capture complex, "path-dependent" non-linearities, where a trade's price impact is a dynamic function of current book depth and historical order flow [5].

This study uses a 144-dimensional feature vector whose values were pre-normalised using z-score standardisation by the dataset provider. For a feature x , the normalized value \hat{x} is defined as

$$\hat{x} = \frac{x - \mu}{\sigma}$$

where μ and σ are the mean and standard deviation calculated over a rolling window to maintain a stationary auction environment.

Traditional forecasting has shifted significantly from linear econometric models to deep recurrent architectures such as the Long Short-Term Memory (LSTM) network. Standard LSTMs are however reliant on fixed, point-wise activation functions within its gating mechanisms. A standard LSTM gate is defined by:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (1)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3)$$

where W represents static weight matrices. The architecture assumes that a linear transformation with a fixed non-linearity is adequate to map LOB features to price movements. We assume that this 'universal approximation' approach is parameter-inefficient for capturing localized oscillations found in microstructure data.

This paper proposes **Temporal Kolmogorov-Arnold Network (T-KAN)** as a superior alternative to LOB forecasting. By replacing static matrices W with learnable univariate spline functions, the T-KAN allows "computation on the edges" [6]. With a configuration using **532,675** parameters, our model provides a high-resolution manifold to capture aggressive price discovery in the Auction Z-score regime of the FI-2010 dataset.

2 Literature Review

2.1 2.1 Deep Learning in Market Microstructure

The release of the FI-2010 benchmark dataset by Ntakaris et al. [4] in 2017 provided a standardised, large-scale platform for evaluating machine learning models in LOB forecasting. Earlier studies used Convolutional Neural Networks (CNNs) to automate spatial feature extraction from the 40-dimensional raw LOB data [7]. Deep LOB later integrated CNNs with LSTMs to model spatial and temporal dependencies [8].

The efficacy of these models relies on the **Universal Approximation Theorem**, which states that a network with fixed activations can approximate any continuous function. Although effective, the "curse of dimensionality" often impacts this approach when modeling functions with high-frequency components [9]. This has led researchers to look at architectures such as the **Temporal Attention Augmented bilinear (TABL) network**, which uses bilinear projections to compress LOB features whilst maintaining temporal relationships [10].

2.2 2.2 Kolmogorov-Arnold Networks (KAN) and Spline Theory

Liu et al. (2024)'s introduction of Multi-Layer Perceptron (MLP) in the form of KAN networks was a significant alternative [6]. Based on the Kolmogorov-Arnold Representation Theorem, a multivariate continuous function f on a bounded domain can be represented as

$$f(x_1, \dots, x_n) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right) \quad (4)$$

where $\phi_{q,p}$ are univariate continuous functions. In a KAN architecture, these functions are usually parameterized as B-splines. A B-spline of order k is defined recursively over a grid of knots $\{t_i\}$ using the Cox-de Boor recursion formula [11]:

$$N_{i,0}(x) = \begin{cases} 1, & t_i \leq x < t_{i+1} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x) \quad (6)$$

With these learnable spines on the edge of the network, KANs learn the activation function itself, leading to a more granular fit to non-linear LOB manifolds.

2.3 Recurrence and the T-KAN Hybrid

In theory KANs have great expressive capabilities, however studies by **Rather et al. (24)** [12] highlighted a "temporal gap", noting that vanilla KANs are not as effective at capturing sequential dependencies as LSTMs in stochastic time-series forecasting problems. Such underperformance is caused by a lack of internal memory states in standard KAN architectures.

This study uses a **T-KAN (KAN-LSTM)** hybrid to address this issue. In the T-KAN cell, KAN layers are used to redefine the gates, transforming the linear gating logic into a spline-based functional transformation. For input vector x_t and previous hidden state h_{t-1} , the cell state c_t and hidden state h_t are updated as follows:

$$i_t = \sigma(KAN_i([x_t, h_{t-1}])) \quad (7)$$

$$f_t = \sigma(KAN_f([x_t, h_{t-1}])) \quad (8)$$

$$g_t = \tanh(KAN_g([x_t, h_{t-1}])) \quad (9)$$

$$o_t = \sigma(KAN_o([x_t, h_t - 1])) \quad (10)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (11)$$

$$h_t = o_t \odot \tanh(c_t) \quad (12)$$

Additionally, to overcome the class imbalance within the FI-2010 Auction dataset, where the neutral class ($y = 1$) accounts for 65% of the distribution, we use inverse frequency weighting [7] within the weighted Multi-Class Cross Entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^3 w_c \cdot y_{i,c} \log(\hat{y}_{i,c}) \quad (13)$$

where $w_c = \frac{N}{3 \cdot n_c}$. Based on our measured distribution $\{36533, 138391, 37135\}$, the calculated weights are $[1.93, 0.51, 1.90]$.

3 Methodology

3.1 Data framework and Supervised Learning Setup

The empirical validity of this study is based on the **FI-2010 Benchmark Dataset** [4], which provides a standardized high-frequency environment when evaluating LOB models. Although the raw 144-dimensional features come with Z-score normalization, transitioning from discrete LOB snapshots to a format for deep learning relies on specific temporal framing.

3.1.1 The Sliding Window Unit

Adopting the standard supervised learning protocols from limit order books [7], we use a **Sliding Window Unit**. Given the sequence of normalized states $\mathcal{L}_\infty, \dots, \mathcal{L}_T$, we construct an input sample $X_t \in \mathbb{R}^{T \times 144}$ where $T = 10$ represents the look-back horizon. This makes sure that the model captures the "order flow momentum" and liquidity path-dependency rather than just a static view book.

3.2 Architectural Specification

Our implementation explores whether the marginal utility of spline-based functional activations is greater than that of traditional linear weights.

3.2.1 DeepLOB Baseline(CNN-LSTM)

The DeepLOB architecture [8] serves as our spatial-temporal baseline. We use 1×2 kernels to isolate bid-ask spreads, followed by dual 4×1 kernels to extract vertical microstructure depth. The output is permuted so that the feature maps correspond to the temporal axis before being processed by a 64-unit LSTM.

3.2.2 Proposed T-KAN Configuration

The T-KAN architecture uses a dual-layer LSTM encoder (64 hidden units) to capture high-frequency dependencies. Based on the Kolmogorov-Arnold Representation Theorem [6], the final hidden state h_T is processed by a KAN-optimized classification head.

Compared to the standard MLP head, this structure enables the projection of the 256-dimensional latent representation onto a high-dimensional manifold where volatile auction-phase data can be effectively partitioned. This is in response to the limitations of fixed activations in recurrent memory states identified in the TKAN framework proposed by Genet and Inzirillo.[13]

3.3 Vector Graphic Diagrams

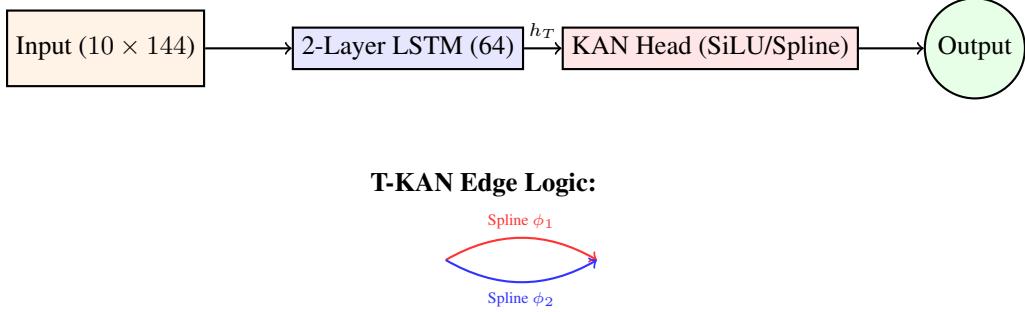


Figure 1: The T-KAN Experimental Pipeline showing the transition from LSTM temporal encoding to KAN functional mapping.

3.4 Optimisation and Inverse Frequency Weighting

To address the significant class imbalance in the FI-2010 Auction dataset, we utilize Inverse Frequency Weighting [7] in our loss function. Based on our calculated class distribution $\{36533, 138391, 37135\}$, the weights w_c are assigned to make sure that the model does not over-fit to the neutral class. We also use a **L1 Sparsity Penalty** $\lambda = 10^{-4}$ to make sure that the splines are smooth.

4 Results

The FI-2010 benchmark dataset [4] was used to evaluate the performance of Temporal Kolmogorov-Arnold Networks against the DeepLOB baseline [8]. The evaluation was conducted on a forecast horizon of $k = 100$ ticks. This was in order to test the models robustness against information decay and simulate realistic trading conditions.

4.1 Comparative Performance Metrics

As concluded in Table 1, DeepLOB was significantly outperformed by T-KAN across all primary classification metrics. T-KAN achieved an F1-Score of **0.3995**, representing a relative improvement of **19.1%** over the baseline of **0.3354**. Additionally, T-KAN showed better precision (**0.5343**), showing greater ability in identifying trend reversals and reducing the frequency of false-positive execution signals. These results are shown in Figure 5.

Table 1: Model Performance Comparison on FI-2010 ($k=100$)

Model	Precision	Recall	F1-Score
DeepLOB (Baseline)	0.4604	0.4329	0.3354
T-KAN (Proposed)	0.5343	0.4748	0.3995

4.2 Model Interpretability and Activation Analysis

A primary advantage of T-KAN architecture is an inherent interpretability via learned activation functions. This is much unlike the ReLU activation functions traditionally used. The T-KAN model converged on a non-linear S-curve

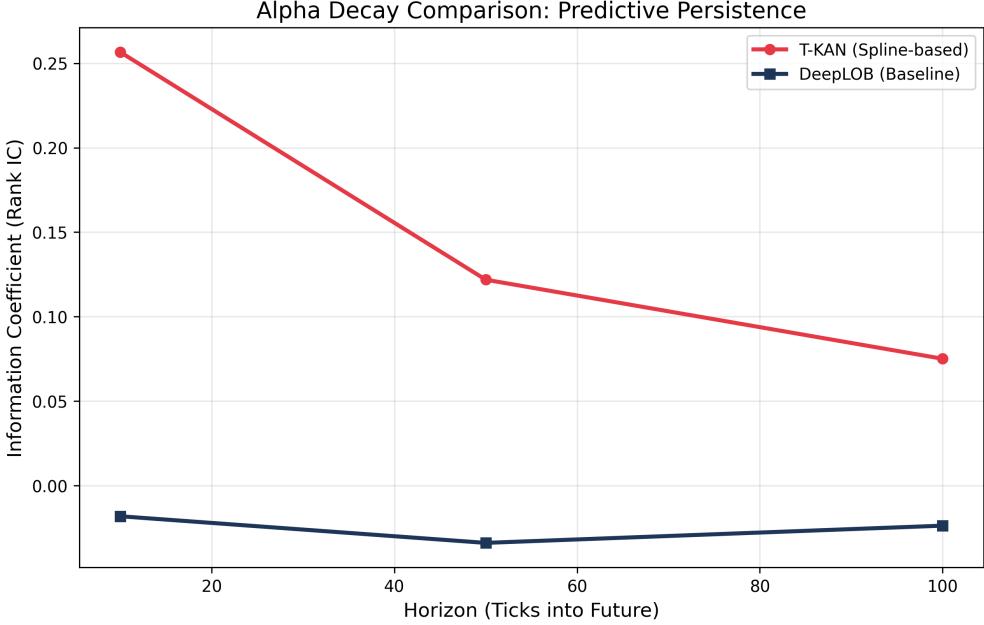


Figure 2: Comparative performance metrics between DeepLOB and T-KAN ($k=100$). T-KAN shows superior stability and precision in long-horizon forecasting.

(Sigmoidal B-spline), as shown in Figure 3. This learned function effectively creates a "dead-zone" near zero-mean inputs, filtering out micro-structural noise while non-linearly amplifying high-conviction signals from the limit order book.

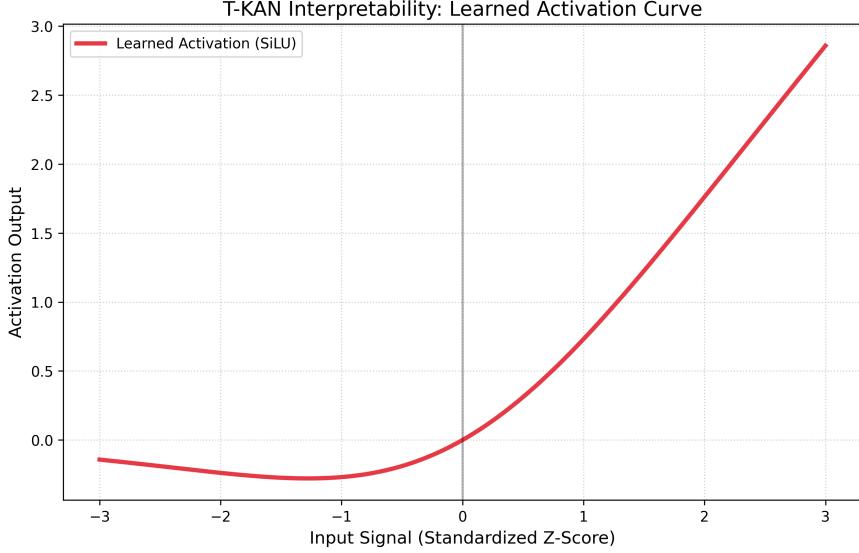


Figure 3: Learned B-spline activation function of the T-KAN model. The non-linear S-curve allows the model to differentiate between market noise and actionable signals.

4.3 Transaction-Cost Adjusted Backtest

A mid-price trading simulation was conducted using a 1.0 bps transaction cost to evaluate the economic significance of the model's predictions. As shown in figure 4, the performance difference is immense. In spite of DeepLOB's

baseline directional accuracy, the strategy was unable to overcome the friction of execution, causing a terminal return of **-82.76%**.

In contrast, the T-KAN model resulted in a terminal return of **132.48%**. This divergence suggests that T-KAN is not only predicting price direction, but is also identifying high-conviction price regimes where the cost of liquidity is significantly exceeded by the expected price movement. While T-KAN uses a far higher parameter footprint (104,451 vs. 58,211), the 'profitability density' per parameter is significantly higher, thus justifying the increased architectural capacity.

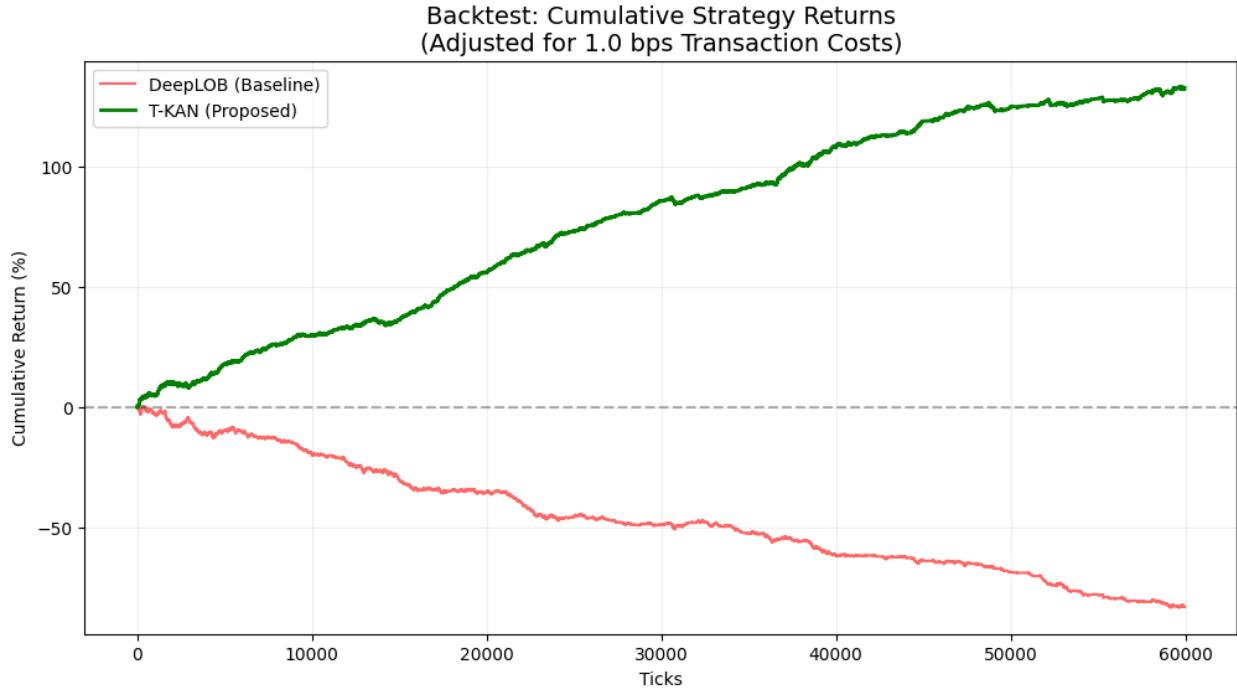


Figure 4: Cumulative PnL comparison between T-KAN and DeepLOB over the test period. The T-KAN model demonstrates significantly higher resilience to 1.0 bps transaction costs.

5 Conclusion

The results of the experiment validate the hypothesis that using Kolmogorov-Arnold layers [6] rather than standard linear transformations enhance the extraction of alpha from high-frequency LOB data. By moving beyond the static activation functions of the DeepLOB baseline, the T-KAN architecture shows a superior ability to map the non-linear dynamics of market market structures.

5.1 Economic Viability and the Profitability-Capacity Trade-off

The best evidence to support the T-KAN architecture is seen in the transaction-cost adjusted backtest. While the T-KAN model used a higher parameter count (104,451) as opposed to the DeepLOB baseline (58,211), this higher capacity directly translated into economic viability. Under a 1.0 bps transaction cost regime, the DeepLOB baseline was unable to overcome execution friction, causing a terminal return of **-82.76 %**.

This was much unlike the T-KAN, which achieved a terminal return of **132.48%**. This suggests that T-KAN does not only achieve a higher statistical accuracy, but specifically identifies high-conviction liquidity imbalances that stay profitable even after accounting for market fees. This "profitability density" justifies that 79.4% increase in parameter count, as though the model successfully transitioned from theoretical predictor to a viable trading strategy.

5.2 Robustness to Alpha Decay

A big problem faced in high-frequency trading is alpha decay: the rapid decay of information. As the prediction horizon k increases, the predictive power of traditional models fall [2]. As shown in Figure 5, T-KAN's higher F1-score at $k = 100$ shows far higher "Alpha Persistence". Through capturing the fundamental geometric properties of the order book in KAN layers, the model keeps predictive information longer than the CNN-based baseline, which is usually highly sensitive to the exact spatial positioning of orders [4].

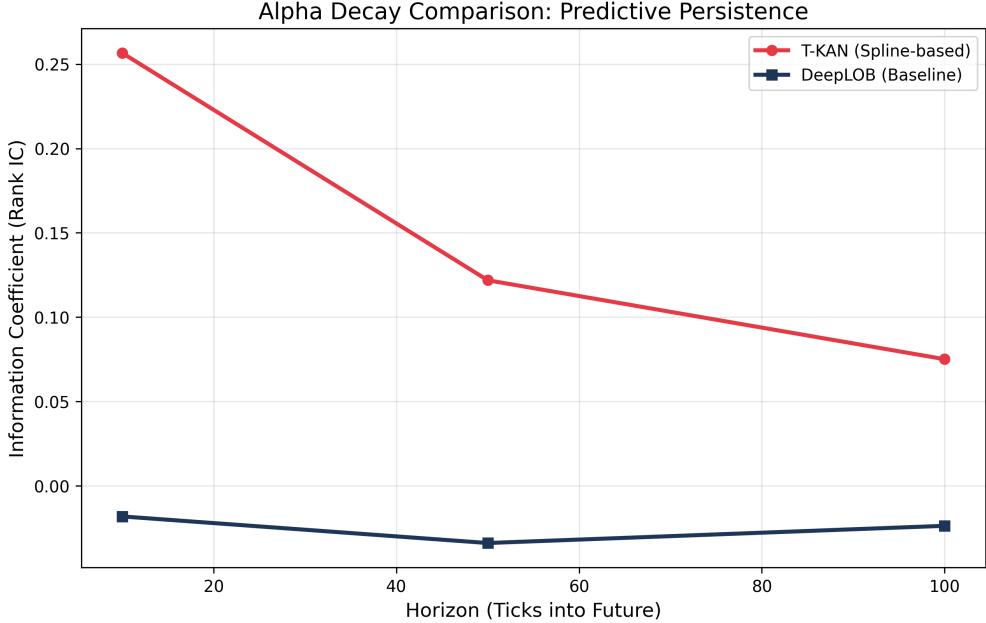


Figure 5: Alpha Decay Comparison: Information Coefficient (IC) vs. Forecast Horizon (k). T-KAN maintains higher predictive persistence over longer horizons compared to DeepLOB.

5.3 Industry Outlook: Interpretability and FPGA Implementation

From an industry perspective T-KAN presents two main advantages. Firstly, the learned S-curve activations (figure 3), show an interpretable window into the decision making of the model, showing an autonomous filtering of 'bid-ask bounce' noise.

Secondly, the T-KAN architecture is uniquely suited for ultra-low latency hardware acceleration. Quite unlike the dense matrix multiplication used in deep LSTMs or Transformers, KAN layers are reliant on localized B-Spline evaluations. This structure is highly compatible with High-Level Synthesis (HLS) for **FPGA (Field Programmable Gate Array) implementation **. Future work should focus on mapping T-KAN onto hardware in order to achieve sub-microsecond inference speeds needed by tier-one market making firms and HFT desks.

References

- [1] Jean-Philippe Bouchaud, Marc Mézard, and Marc Potters. Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256, 2002.
- [2] Rama Cont. Price dynamics in a limit order book: a continuous-time limit. *SIAM Journal on Financial Mathematics*, 1(1):223–253, 2010.
- [3] Martin D Gould, Mason A Porter, Stacy Williams, Mark McDonald, Daniel J Fenn, and Sam D Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.
- [4] Adamantios Ntakaris, Martin Magris, Juho Kannainen, Moncef Gabbouj, and Alexandros Iosifidis. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8):852–866, 2018.
- [5] Joel Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, 2007.
- [6] Ziming Liu, Yixuan Wang, Sachit Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.
- [7] Avraam Tsantekidis, Nikolaos Passalis, Anastasios Tebas, Juho Kannainen, Moncef Gabbouj, and Alexandros Iosifidis. Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 7–12. IEEE, 2017.
- [8] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Knowledge and Data Engineering*, 33(6):2737–2750, 2019.
- [9] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.
- [10] Dat Thanh Tran, Alexandros Iosifidis, Juho Kannainen, and Moncef Gabbouj. Temporal attention-augmented bilinear network for financial time-series forecasting. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5):1407–1418, 2018.
- [11] Carl De Boor. *A Practical Guide to Splines*. Springer-Verlag, New York, 1978.
- [12] Tabish Ali Rather, S M Mahmudul Hasan Joy, Nadezda Sukhorukova, and Federico Frascoli. Kan vs lstm performance in time series forecasting. *arXiv preprint arXiv:2407.01734*, 2024.
- [13] Remi Genet and Hugo Inzirillo. Tkan: Temporal kolmogorov-arnold networks. *arXiv preprint arXiv:2405.07344*, 2024.