IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 22 June 2023, accepted 2 July 2023, date of publication 7 July 2023, date of current version 17 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293064

RESEARCH ARTICLE

A Signature Transform of Limit Order Book Data for Stock Price Prediction

THENDO SIDOGI[1], WILSON TSAKANE MONGWE[1], RENDANI MBUVHA[2,3], PETER OLUKANMI[1], AND TSHILIDZI MARWALA[4], (Senior Member, IEEE)
[1]School of Electrical Engineering, University of Johannesburg, Auckland Park 2000, South Africa
[2]School of Statistics and Actuarial Science, University of Witwatersrand Johannesburg, Johannesburg 2006, South Africa
[3]School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K.
[4]United Nations University, Tokyo 150-8925, Japan

Corresponding author: Thendo Sidogi (thendos@gmail.com)

ABSTRACT A novel approach is presented for predicting the mean-mid stock price by utilizing high-frequency and complex limit order book (LOB) data as inputs for machine learning algorithms. Specifically, the proposed approach uses rough path theory to extract signature path features from the LOB data and compresses them for training machine learning models. We compare the performance of this approach to standard autoregression methods and demonstrate its superior performance in terms of model prediction error. In addition, we use deep neural networks (DNN) and random forest (RF) to further test the proposed approach's performance compared to standard approaches using raw LOB data. The findings show that Sig-DNN, DNN with signature features, outperforms DNN with raw LOB data in terms of prediction error and efficiency, while Sig-RF, RF with signature features, underperformed RF with raw LOB data in terms of prediction but not in efficiency. We evaluate the proposed approach on multiple exchanges, including the Johannesburg and New York Stock Exchanges, and the results demonstrate better prediction error and efficiency performance in developed markets compared to emerging markets. Overall, the study highlights the potential of using signature path features with machine learning techniques for predicting the mean-mid stock price.

INDEX TERMS Limit order book, rough paths theory, signatures, DNN, RF, regression, mean-mid price, machine learning.

I. INTRODUCTION

In a trade life cycle, quantitative portfolio management encompasses trade execution to implement investment strategies into practice. The process of trade execution generates a vast amount of data, including high-frequency and high-dimensional Limit Order Book (LOB) data. These data comprise various market events such as order placement and cancellation, which are essential components of trade execution in many exchange trading platforms [1]. However, understanding the underlying dynamics and market phenomena driving this data can be challenging [2]. To address this challenge, buy-side quantitative analysts seek methods to compress high-frequency data streams without losing essential information for stock price prediction [3]. Although stock price prediction can be seen as a traditional time series prediction task, high-frequency data presents a unique challenge due to issues such as data frequency and volume. Practitioners aim to overcome this challenge in order to gain or maintain a competitive advantage over their peers [4], [5]. In this study, we address the challenge and present an approach to predict the mean-mid price movement using data stream features derived from the path signature of sampled LOB data paths.

The frequency and volume of order books were comparatively lower in earlier years. In the literature, we came across the application of regression analysis where the authors studied the relationship between trading activity and order flow dynamics in LOB from the Swiss exchange, utilizing a Probit model [6]. Similarly, Yu [7] tries to extract information

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

from order information and submission based on the ordered Probit model. The data used in the study was from the Shanghai Exchange, which was considered an emerging market exchange at the time.

Cao et al. [8] examines the depth of different levels of an order book from the Australian Exchange by using an Auto Regression (AR) model of order 5. They found that levels beyond the best bid and best ask prices provide moderate information regarding the true value of an asset. Similarly, Cenesizoglu et al. [9] use a linear vector AR system for price prediction and found that LOB help predict short-run mid-quote return. The LOB data was from the German stock exchange, a developed market exchange. For price jump prediction of intraday stock prices using LOB features, the research by Zheng et al. [10] used logistic regression, where the experiment was based on the forty largest French stocks of CAC40 index. Liu et al. [11] developed a multivariate linear model to explain short-term stock price movement on LOB's from the US stock market where a bid-ask spread is used for classification purposes. The results from the regression models showed improved performance in explaining short-term stock price movement using LOB data.

Over the years, the number of trades within the most prominent exchanges has increased dramatically since the advent of computer-based trading [12], [13]. Orders could be submitted or modified within milliseconds, or even faster during specific time intervals. High Frequency Trading (HFT) strategies were starting to be deployed more within exchanges for market marking purposes [14]. HFT is a type of trading strategy where traders execute many trades and generate small profits for small changes in stock prices in each transaction. This resulted in a replacement of traditional quote-driven markets with order-driven trading platforms [15]. The high-frequency nature of the strategy results in many chaotic data points [16]. While it improves efficient price discovery, it creates the challenge of generating high-dimensional data. This challenge offers a suitable framework for applying machine learning techniques in dealing with big data. Tran et al. [17] used a tensor framework to learn multi-dimensional data for HFT trading.

A comparison is performed in [18] on state-of-the-art models with different prediction horizons. It was found that deep neural networks outperforms the state-of-the-art CNN-LSTM architectures in different time horizons measured various performance metrics including Balanced Accuracy (for imbalanced datasets), weighted F-score and Matthews Correlation Coefficient. Ntakaris et al. [19] introduced a benchmark data set and used different machine learning methods for mid-price prediction. Nousi et al. [20] used a feature engineering approach on LOB to predict stock price of future movement using machine learning algorithms. Similar to Passalis et al. [21], their feature engineering technique included handcrafted features related to rolling moments of the path and machine learning inspired features from autoencorders and bag-of-features models. The previous studies are relevant to this research as they utilize feature engineering techniques to predict mid-price using machine learning algorithms. However, this study proposes a different approach and employs rough path theory to generate signature features from LOB data.

Hambly and Lyons [22] established the notion of a signature as an informative transformation of a multidimensional time series. A signature is a top down description for unparameterized path that describes a path segment through its effects of a highly oscillatory and non-linear system. These systems include the financial market system and interactions within market microstructures. One possibility that Ni et al. [23] introduced is to use the concept for understanding financial data, and they have emphasized the potential of this approach for machine learning and prediction. We refined the theoretical suggestions and evaluated them against real-world financial data, including LOB data for different stocks and exchanges. In each experiment discussed later, we explore different dimensional parts of the signature, referred to as signature order or truncation, that capture essential information for characterizing various features.

The theory of rough paths was developed in the 90s by Lyons [24]. Since then, the field of study has evolved considerably and has been applied in different areas. Rough path theory is well positioned for machine learning and time series application due to its ability to extract top-down-level features for well-defined paths or data streams. Notable areas of application include synthetic financial data generation [25], handwriting recognition [26], sepsis detection [27], human action recognition [28], drone identification and analysing energy usage data [29].

In the literature, we find little research on using rough path theory to generate signature features of LOB for stock price prediction. Although rough path theory was applied to high-frequency financial tick data, this was only to show the ability of the method to compress and capture sufficient information in unknown trading strategies involving this data [30]. We aim to provide a framework for dealing with multi-dimensional high-frequency LOB data. This study proposes a novel solution utilizing rough path theory to create signature path features. The signatures coefficients are then used as input features for machine learning algorithms to predict the movement of stock mean-mid prices. It is also the first study to compare the effectiveness of this approach in different types of markets, including an African stock exchange.

Contributions: Our contributions in this work can be summarised as follows:

- We propose a first in-literature use of rough path theory to generate price path features on LOB data for mean-mid price prediction. The proposed method effectively compresses high-frequency LOB data and generates signatures for machine learning algorithms to predict stock's mean-mid prices.
- Second, In this study, we present the first in literature comparison of LOB data from both emerging and developed markets in order to understand the potential of

LOB data as an alternative data source for stock price prediction in different market conditions.

The remainder of this paper is organized as follows: In section II, we discuss the definition of signature of a path and its properties and apply the definition to the data used in this study. We discuss the proposed experiment of the study and look at snapshots of the LOB data used in section III. Section IV provides a description of the machine learning algorithms used for prediction. We then present the results, including plots of model prediction and table of model performance for all the machine learning models in section V. Section VI concludes the paper and outlines ideas for future research.

II. SIGNATURE OF A PATH

The application of mathematical finance within the finance industry has been concerned with estimating future real-world events based on the past sequence of events or a well-defined path established by the series of events [30]. We want to understand a stream's effects on a complex system, such as the decision to buy or sell a stock, using a function. One possible example of the function f is a stock price at time t. Such non-smooth paths may represent sequential data or time series, typically consisting of successive measurements made over a time interval. Using these data streams, quantitative analysts will decide how to summarise information about the paths using specific parameters. These can be drift components on the path or some volatility measure of the path. For example, estimating mean stock returns and standard deviation to construct a minimum variance portfolio. However, studies show that these parameter estimates are based on specific model assumptions on the dynamic of ever-changing phenomena in the financial markets [25], [31].

This study aims to establish how the evolving nature and well-defined path of LOB data affect the non-linear system of stock price prediction. This can be achieved using the controlled differential equation below

$$dY_t = f(Y_t)dX_t \tag{1}$$

where X_t is the input or driver path to the system. Here, X_t can be seen as a sequence of events within the LOB data. Y_t is the response to the input path. In our case, Y_t can be seen as a stock price prediction or a buy or sell signal. Solving for equation (1) can be challenging since we are trying to determine how small increments in X affected small increments in Y. A popular way to do it is often using a parameterized algebraic approach of fitting a polynomial function. The challenge, however, is real-world paths or data streams over time are often random, very noisy, unparameterized and non-analytic. For a function f and path Y, we say that Y solves equation (1) when the equation (2) is met

$$Y_t = y_0 + \int_0^t f(Y_t)dX_t \tag{2}$$

Rough paths theory suggests that there exist unparameterized paths describing data streams. Rough path theory is a mathematical framework that allows for the analysis of irregular and non-smooth data streams. It provides a way to extract meaningful information from these paths by representing them as sequences of iterated integrals.

It is found in [24] that signature of a path X_t is the solution to equation (1) and solves for equation (2). These are sequential Picard approximations to the true solution for equation (2). It is a faithful invariant and top-down description of unparameterized paths generating the best coefficients describing the path. The resulting output transforms into tensor algebra when translating paths or data streams to signatures. The tensor product of a path in d dimension is defined below in equation (3)

$$T(\mathbb{R}^d) = \bigoplus_{k=0}^{\infty} \left(\mathbb{R}^d\right)^{\otimes k} \tag{3}$$

It contains different information sets at different levels or truncations (since we may not necessarily reach infinity during computation). Each level or integral grouping describes an input path's extra dimension of variability.

We define the signature of a path as follows. Set X as a continuous function with finite length $[0; T]$. The signature of a continuous path X_t over an interval $a \leqslant t \leqslant b$ denoted by $S(X_t)_{a,b}$ are the elements of $T(\mathbb{R}^d)$ defined by the following sequence of real numbers:

$$\begin{aligned}
S^{(0)} &= 1,\\
S_i^{(1)} &= \int_{a<v<b} dX_v^i \qquad (1 \leqslant i \leqslant d),\\
S_{i,j}^{(2)} &= \int_{a<u<v<b} dX_u^i dX_v^j \qquad (1 \leqslant i,j \leqslant d),\\
&\ \vdots\\
S_{i_1,\ldots,i_k}^{(k)} &= \int_{0<u_1..<u_k<T} dX_{u_1}^{i_1} \ldots .dX_{u_k}^{i_k}\\
&\quad (1 \leqslant i_1, \ldots, i_k \leqslant d),\\
&\ \vdots
\end{aligned}$$

where integration limits correspond to integration over a simplex in a higher dimension. We note that it may be impractical to compute the full signature of a path so the truncated signature of order k, $S(X_t)_{a,b}^{(k)}$, is used.

We highlight some of the properties below. Further proofs can be found in [22] and [32].

1) **Invariant with time reparametrisation**: for continuous and non-decreasing function $\gamma : [0, T] \longmapsto [S, U]$ and $0 \leqslant s \leqslant t \leqslant T$. Then path integrals (hence signatures) are invariant under time reparametrisation of γ given by:

$$\begin{aligned}
&\int_{s<u_1..<u_k<t} dX_{u_1}^{i_1} \ldots dX_{u_k}^{i_k} =\\
&\qquad \int_{\gamma(s)<u_1..<u_k<\gamma(t)} dX_{\gamma^{-1}(u_1)}^{i_1} \ldots dX_{\gamma^{-1}(u_k)}^{i_k}
\end{aligned} \tag{4}$$

2) **Multiplicative and concatenation property**: there exists a binary operation $\otimes$, such that if any $a < u < b$ then

$$S(X_t)_{a,u} \otimes S(X_t)_{u,b} = S(X_t)_{a,b} \tag{5}$$

where the binary operation is discussed in [24].

Since financial time series data is not in continuous trajectories, for instance, the market only shows the prices of the securities at specific times. Therefore, we use the path signature's first property 4 to transform our stream dataset while ensuring invariance under time reparametrisation.

A. SIGNATURE OF LOB DATA STREAM

In our previous definition of a path signature, we focused on continuous data paths. However, many data streams, such as LOB data, evolve over time without being continuous. LOB data, for example, represents market orders placed at various times based on participants' discretion. To capture the dynamics of such data streams, we need to define the path signature using a signature of stream data, introducing continuity. The path signature, computed by iteratively integrating and accumulating the path over different time intervals, represents higher-order features that capture the cumulative information within the data. This approach effectively compresses the data stream while preserving its essential characteristics. By applying rough path theory and extracting signature path features, we can summarize the complex and high-dimensional LOB data, making it suitable for further analysis and prediction tasks.

Levin et al. [23] defined the signature of data streams by constructing axis paths through rectilinear interpolation instead of piece-wise linear interpolation. Another way to transform to a continuous path is using lead-lag transformation, which maps a one-dimensional path into a two-dimensional path. In particular and similar to work by Gyurko et al. [3], given a stream of LOB data $(X_{t_i})_{i=0}^{N}$ in $\mathbb{R}^d$, we define its lead-transformed stream $(X_j^{lead})_{j=0}^{2N}$ by

$$X_j^{lead} = \begin{cases} X_{t_i} & \text{if } j = 2i \\ X_{t_i} & \text{if } j = 2i - 1 \end{cases}$$

and its lag-transformed stream $(X_j^{lag})_{j=0}^{2N}$ is defined as

$$X_j^{lag} = \begin{cases} X_{t_i} & \text{if } j = 2i \\ X_{t_i} & \text{if } j = 2i + 1 \end{cases}$$

the lead-lag-transformed stream takes values in 2-dimension and is defined by the paired stream

$$(X_j^{lead-lag})_{j=0}^{2N} = (X_j^{lead}, X_j^{lag})_{j=0}^{2N}$$

Hence, it can can be seen that the signature of the lead dimension can be written as

$$S_{t_0,t_N}(X) = S_{0,2N}(X^{lead})$$

and similarly the signature of the lag dimension

$$S_{t_0,t_N}(X) = S_{0,2N}(X^{lag})$$

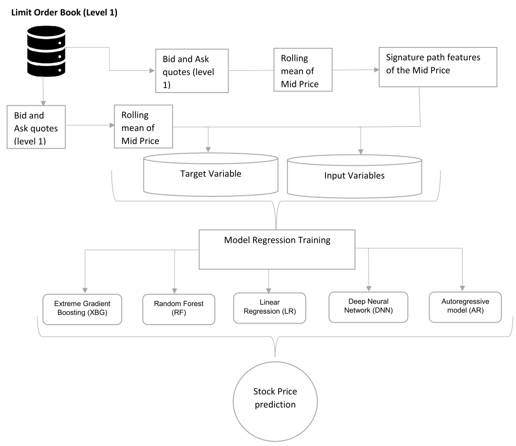


FIGURE 1. Proposed framework for mean-mid stock price prediction using signature path features.

We can therefore transform any stream of data to continuous data and still be able to capture for example the quadratic variation of high frequency LOB data streams using lead-lag transformation.

III. DATA AND RESEARCH DESIGN

We designed our experiment to see if we can use path signatures as features to machine learning models in predicting stock price movement in LOB data. In other words, if X is a stream of data from an electronic LOB as a sequence of prices, and the effects of this stream are given by a functional f that transforms the streams of prices to a prediction made by a investor given by $Y = f(X)$, our task is to predict $f(X)$ given any X based on the data in the observed learned LOB data. We then compare the results to the benchmark AR model.

In summary, our proposed workflow is depicted in Figure 1 shown below.

A. DATA STREAM

We start by defining some LOB features. More references and information about the topic can be found in [33] and [34]. Our experiment is based on LOB data on two stocks in two different exchange markets: AMAZON stock in the New York Stock Exchange (NYSE) and NASPERS in the Johannesburg Stock Exchange (JSE). It is important to note that JSE and NYSE are very different in size and trading volumes since the JSE is from emerging market economy and NYSE is from a developed market economy. Exchanges from developed markets have far greater number of listed companies and have higher trading volumes compared to emerging market exchanges. LOB consists of different levels indicating trades submitted by market participants. The number of levels can vary depending on the exchange and liquidity, ranging from level 1 to level 10. These levels provide varying degrees of information on how market participants interact, with higher levels containing more comprehensive information on the most recently executed trade. However, for this study, we only consider level 1 LOB data.

TABLE 1. **First 10 trades of AMAZON stocks from LOB on NYSE on 2012-06-21.**

AMAZON LOB DATA SAMPLE			
Time (hh:mm:ss:000)	Bid Price	Ask Price	Mid Price
09:30:00.000	223.18	223.82	223.565
09:30:00.762	223.75	223.95	223.85
09:30:00.812	223.75	223.95	223.85
09:30:00.856	223.84	223.95	223.895
09:30:00.857	223.75	223.95	223.85
09:30:00.864	223.75	223.95	223.85
09:30:00.864	223.84	223.95	223.895
09:30:00.874	223.84	223.95	223.895
09:30:01.273	223.84	223.86	223.85
09:30:01.274	223.84	223.95	223.895

TABLE 2. **First 10 trades of NASPERS stocks from LOB on JSE on 2021-10-26.**

NASPERS LOB DATA SAMPLE			
Date-Time	Bid Price	Ask Price	Mid Price
2021-10-26T06:32:50.788085406Z	2598.00	2590.00	2594.00
2021-10-26T06:32:50.963285119Z	2598.00	2590.00	2594.00
2021-10-26T06:32:50.963285119Z	2598.00	2590.00	2594.00
2021-10-26T06:32:50.963285119Z	2598.00	2590.00	2594.00
2021-10-26T06:32:55.495386302Z	2598.00	2551.34	2574.67
2021-10-26T06:32:55.675600741Z	2598.00	2551.34	2574.67
2021-10-26T06:34:31.131275605Z	2598.00	2551.34	2574.67
2021-10-26T06:36:44.519316467Z	2598.00	2551.34	2574.67
2021-10-26T06:36:44.519316467Z	2598.00	2551.34	2574.67
2021-10-26T06:36:44.519316467Z	2598.00	2551.34	2574.67

The data contains the first 4000 time ticks of a single trading day for both AMAZON and NASPERS with an 80:20 training testing split. The dates selected for AMAZON and NASPERS are 2012-06-21 and 2021-10-26 respectively. For every tick, the LOB data contains the following:

- Bid price: the highest price submitted electronically by a buyer to pay for a specified number of stocks at any given time.
- Ask price: the lowest price submitted electronically by a seller to offload a specified number of stocks at any given time.
- Mid price: the average of Bid and Ask price

Table 1 and 2 show that the electronic LOB data is a discrete data stream. We also indicate the five-number summary of the mid-price from each dataset in Table 3. It is easy to notice that numerically, the magnitude of the difference in maximum and minimum mid-price is small in each data stream. This is the nature of LOB data where there is a small price difference in successive trades.

As mentioned above, signatures are iterated integral on continuous paths. Therefore, we need to transform the LOB data stream to a continuous stream using the lead-lag transformation described above. Figure 3 and 2 plots the transformation from AMAZON mean-mid price stream to continuous data in the lead-lag space.

B. MACHINE LEARNING MODEL BASED ON SIGNATURES

As discussed earlier, several machine learning models have been used on high-frequency LOB. However, we do not know of any papers using signature features for stock price prediction. In this study, we propose the use of signature features on LOB data for stock price prediction using machine learning models. Signatures, in the context of machine learning algorithms, are higher-order features that are derived from the paths or trajectories of data. They capture important information about the cumulative behaviour and dependencies within the data.

TABLE 3. **Five number summary for the LOB datasets.**

Statistic Summary	AMAZON Mid Price ($)	NASPERS Mid Price (R)
Minimum value	220.520	2573.870
25th Percentile	221.265	2623.333
50th Percentile	222.655	2637.155
75th Percentile	223.875	2641.225
Maximum Value	226.025	2659.9

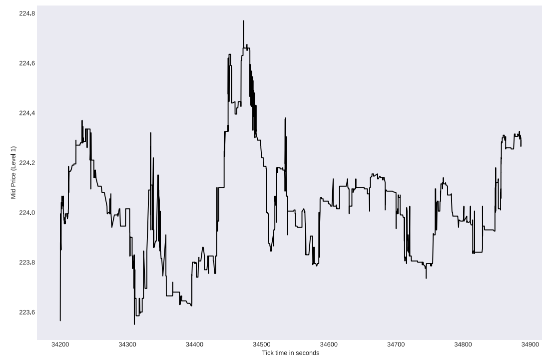


FIGURE 2. **Mean-mid price path stream data.**

The definition of the signature indicates that functions on paths are equal to linear functions on signatures, implying that information about paths can be obtained from their signatures. Levin et al. [23] shows the proof of the theorem. This result implies that the desired function can be accurately approximated by linear functions, including linear regression and other machine learning algorithms. As a result, the approach in this chapter for solving equation (1) can be summarized as follows:

- Given a training sample of a rolling window of size 30 ticks of LOB data $\{(X_t, Y_t)\}_{i=0}^{N}$ where Y_t is the target prediction of mean-mid price and X_t is the stream of data in Table 1 and 2.
- The target prediction in the experiment is the average of 5 consecutive mid-price ticks starting on the 31st.
- Transform rolling window of stream data to a continuous path using lead-lag transformation in the form while including lead time variable z_j^{lead}

$$(\hat{X}_j)_{j=0}^{2N} = (z_j^{lead}, X_j^{lead}, X_j^{lag})_{j=0}^{2N}$$

- Compute $S(\hat{X}_j)^{(n)}$, the truncated signature of order n over a specific interval using Esig python package.
- Apply various machine learning algorithms with appropriate architecture against the truncated signature to solve the controlled differential equation.

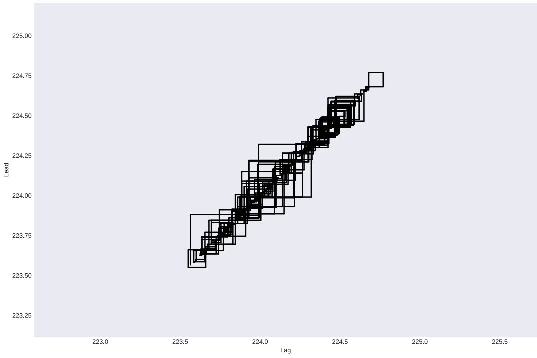


FIGURE 3. **Lead-lag transformation of the financial data stream.**

A signature's dimension depends on the extent of the underlying space and the level at which the signature is truncated. The depth of the signature does not depend on the number of samples or increments used to compute it.

C. PERFORMANCE METRICS

To evaluate the performance of the proposed algorithm, we employed specific performance metrics. The predictive accuracy of the models for the mean-mid price was assessed using root mean square error (RMSE) and mean absolute error (MAE) metrics. Additionally, we utilized the Diebold-Mariano test [35] to determine if there is a statistically significant difference between the forecasts generated by the two different models. In the experiment, we will compare the performance of the proposed model to that of the AR model as the benchmark. While the test does not indicate which forecast is superior, it provides a p-value that helps determine if there is a notable distinction between the two forecasts.

IV. MACHINE LEARNING ALGORITHMS

The signature is a reliable transform of a multidimensional data stream. Levin et al. [23] suggested its application on financial data streams and emphasized its potential for machine learning and prediction. This section discusses the machine learning models utilized for prediction.

A. RANDOM FOREST (RF)

Random forest is a machine learning algorithm that uses an ensemble model that combines several base decision trees to produce an optimal predictive model and make point or classification predictions [36]. Decision trees are models with a high variance and low bias. When increasing the number of decision trees, the random forest reduces the high variance by having uncorrelated individual decision trees and taking the average class prediction from each decision tree while the bias remains constant. In each training step, samples from the training sets are selected randomly with replacements and a subset of features is selected randomly and whichever feature provides the best split or decision is used.

In our study, random forest is a regressor consisting of an ensemble of decision tree-structured regressors $\{h(x, \theta_k), k = 1, \ldots\}$ where $\{\theta_k\}$ are independent identically distributed random vectors and each decision tree predicts the mean of mid-price using path signature features as inputs. The number of decision trees to be boosted was 1100, with a maximum depth of 2 while minimizing the mean square error.

B. EXTREME GRADIENT BOOSTING (XGB)

State-of-art machine learning techniques for classification or regression tasks include extreme gradient boosting, first introduced by Friedman [37]. Gradient boosting is an optimisation problem, where the goal is to locate the minimum or minimise the model's loss function by adding weak learners using gradient descent. During the learning process, decision trees are added at each stage, and the existing trees in the ensemble are not replaced. The contribution of the weak learner to the ensemble is based on the gradient descent optimisation process performed on each weak learner. The calculated contribution of each tree is then based on minimising the overall error of the composite learner.

In this study, the XGB is a regressor with the inputs x_i^p and y_i target set can be defined as $\{(x_i^p, y_i)\}_{i=1}^n$. We use a differentiable function $L(y, F(x))$, where $F(x)$ is the function mapping x^p to y and is the weaker model within the ensemble tree with a number of iterations m. The number of decision trees to be boosted was 100 with a maximum depth of 3 for each tree.

$$\gamma_{jm} = \underset{\gamma}{\operatorname{argmin}} \sum_{x_i \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma) \tag{6}$$

$$F_m(x) = F_{m-1} + \sum_{j=1}^{J_m} \gamma_{jm} 1_{R_{jm}}(x) \tag{7}$$

where chooses a separate optimal value γ_{jm} for each of the tree's regions, instead of a single γ_m.

The XGB model has the benefit of better interpretability in which we can estimate the relative importance of input features in predicting stock price directional movement. The number of decision trees to be boosted was 950 with a maximum depth of 4 for each tree.

C. DEEP NEURAL NETWORKS (DNN)

We also employ a DNN to predict the mean of stocks' mean-mid price. It has three layers of interconnected neuron units, through which the data are transformed. The number of neurons on the input layer depends on signature truncation, where the size will always match the number of coefficients describing the path. The hidden layer has a ReLU activation function, whereas the output layer uses a linear function. The hidden layer has a dimension of 256, with a total of 33024 parameters. The last output layer with the linear function has a total of 257 parameters. The Adam optimiser was used to update all the trainable parameters. The objective function is the mean absolute error loss function that reduces the uncertainty in the distribution of prediction error.

D. LINEAR REGRESSION (LR)

In this study, we also use the linear regression model. In the experiment, the linear regression model will describe the relationship between the mean-mid stocks price prediction Y based on one or more independent variables, X which are signature path features. The dependent variable is also called the response variable. We modelled the mean-mid stock price using signature features generated as independent variables from the observed path using equation (8).

$$Y_t(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \tag{8}$$

where Y is the predicted mean-min stock price, β_i is the parameter for $i - th$ coefficient of the signature x_i and m is the number of signature coefficients. In this study, m depends on the signature truncation.

E. AUTO REGRESSION (AR)

We utilise AR as a benchmark for the proposed models for the experiment. AR is a time series analysis method used to model data that exhibits correlation with its own past values. It assumes that the current value of a variable depends linearly on its previous values, with the degree of dependence determined by a parameter called the autoregressive coefficient. The standard parameter for AR is denoted as p, representing the number of lagged values of the variable used to predict the current value. The AR model of order p is expressed as

$$Y_t = \sum_i (\eta_i * Y_{t-1}) + c + \epsilon_t \tag{9}$$

where Y_t is the current mean-mid stock price, c is a constant, η_i represents the auto regressive coefficients for $i = 1, 2, \ldots, p$, and ϵ_t is the error term. The choice of p will depend on the signature truncation.

V. EXPERIMENT RESULTS

Figures 4-7 show the mean-mid price prediction results of AMAZON and NASPERS by the best-performing machine learning model using their signatures from LOB data streams at different truncations. We used DNN and RF for NASPERS and AMAZON respectively. The Figures demonstrate that AR tends to produce more linear plots, whereas the proposed use of rough path theory on LOB is more effective at capturing the non-linear relationships and patterns inherent in the data.

A. PLOTS OF THE TRAINING AND TESTING EXPERIMENT RESULTS

Figures 4 (a) - 7 (b) show plots of the training experiment results at different signature orders. Our findings reveal that the proposed machine learning models tend to over-fit during training, with this effect being more pronounced for the AMAZON stock price data. We think over-fitting is exacerbated because the RF model used is highly flexible and when the training data may contain a lot of redundant information because of the nature of LOB data.

TABLE 4. p-value from Diebold-Mariano Test.

AMAZON FORECAST				
Signature/ lagged order	XGB	RF	LR	DNN
2	0.0290	0.0610	0.7642	0.9497
3	0.1960	0.0438	0.6838	0.8668
4	0.0294	0.3311	0.0819	0.0544
5	0.3786	0.9374	0.1221	0.1793

Figures 4 (c) - 7 (d) show plots of the testing experiment results. It is visible to notice that the red line tends to move with the green line but not with the blue line. This indicates a better model fit using signature path features compared to using lagged features in the AR model. This is because signature path features contain the inputs path's extra dimension of variability when compared to lagged features. It is further supported by the observation that AR tends to produce more linear plots, while machine learning models incorporating signature path features are more adept at capturing the non-linear relationships and patterns inherent in the data. In Figure 7 (b) and (d), we observe a significant decline in performance compared to lower signature truncations. We think this is because of the factorial increase in the number of significant features which may require a change in the deep neural network architecture.

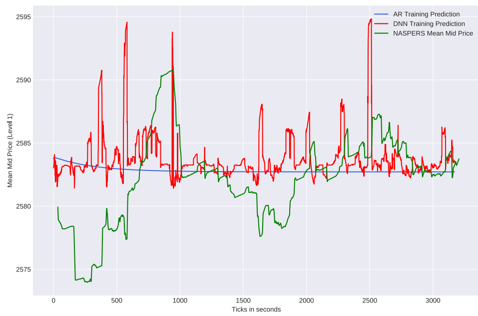
Table 4-5 displays the p-values from the Diebold-Mariano test, which compares the forecasts generated by the XGB, RF, LR, and DNN models with the forecasts from the AR benchmark. The p-values indicate the statistical significance of the difference between the two sets of forecasts.

Table 4 shows the results which represent different signature orders, the XGB and RF models consistently show low p-values across all lag orders, indicating that their forecasts are significantly different from the AR benchmark. The LR and DNN models, on the other hand, have higher p-values, suggesting that their forecasts are not significantly different from the AR benchmark. Overall, the results from Table 4 suggest that the XGB and RF models outperform the LR and DNN models in terms of generating significantly different forecasts compared to the AR benchmark model.

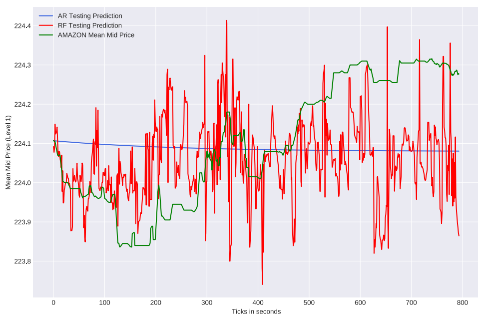
In Table 5, the results show that all the machine learning models (XGB, RF, LR, and DNN) consistently exhibit very low p-values across different lag orders. This suggests that the forecasts generated by these models are significantly different from the forecasts of the AR benchmark model. These findings indicate that the machine learning models outperform the AR model in terms of generating significantly different forecasts compared to the AR benchmark model. However, it is important to note that the p-values do not provide information about the relative accuracy or superiority of the models in terms of forecast performance. To assess the relative accuracy performance, we will measure the MAE and RSME.

Tables 6-9 show the results. There were significantly higher model prediction errors on the NASPERS LOB data than the AMAZON data across all performance

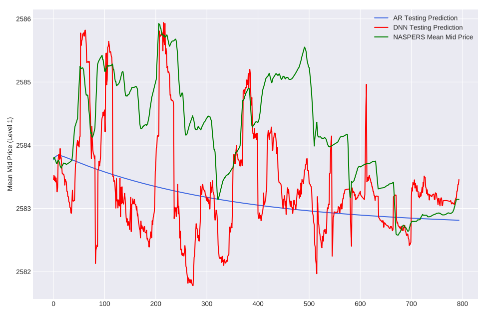
(a) AMAZON mean-mid price path training experiment data.



(b) NASPERS mean-mid price path training experiment data.



(c) AMAZON mean-mid price path testing experiment data.

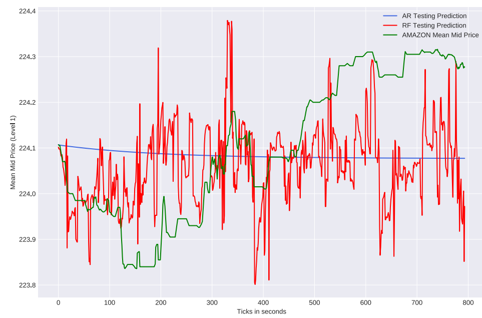


(d) NASPERS mean-mid price path testing experiment data.

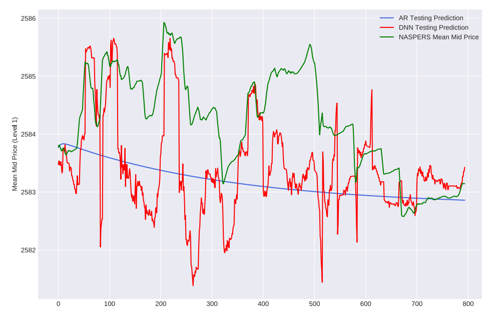
FIGURE 4. Mean-mid price prediction using signatures of order $n = 2$ from LOB data stream.

(a) AMAZON mean-mid price path training experiment data.

(b) NASPERS mean-mid price path training experiment data.



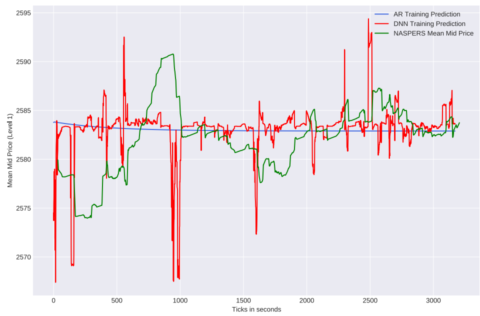
(c) AMAZON mean-mid price path testing experiment data.



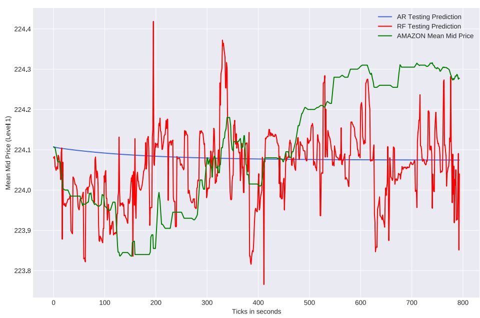
(d) NASPERS mean-mid price path testing experiment data.

FIGURE 5. Mean-mid price prediction using signatures of order $n = 3$ from LOB data stream.

(a) AMAZON mean-mid price path training experiment data.



(b) NASPERS mean-mid price path training experiment data.

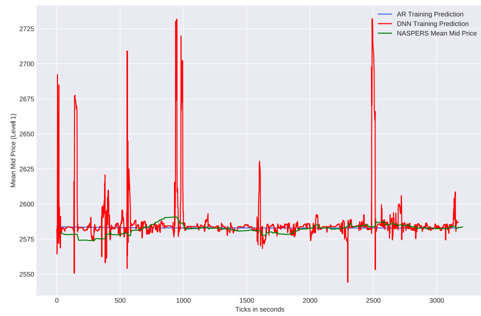


(c) AMAZON mean-mid price path testing experiment data.

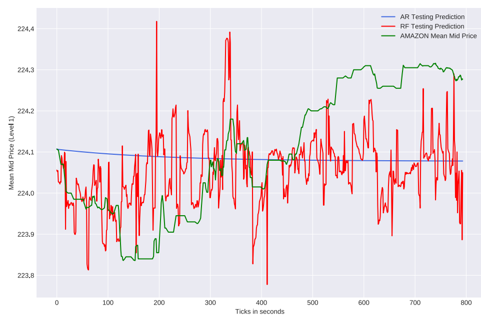
(d) NASPERS mean-mid price path testing experiment data.

FIGURE 6. Mean-mid price prediction using signatures of order $n = 4$ from LOB data stream.

(a) AMAZON mean-mid price path training experiment data.



(b) NASPERS mean-mid price path training experiment data.



(c) AMAZON mean-mid price path testing experiment data.

(d) NASPERS mean-mid price path testing experiment data.

FIGURE 7. Mean-mid price prediction using signatures of order $n = 5$ from LOB data stream.

TABLE 5. **p-value from Diebold-Mariano Test.**

NASPERS FORECAST				
Signature/ lagged order	XGB	RF	LR	DNN
2	0.0057	0.0007	0.0001	0.0001
3	0.0026	0.0078	0.0003	0.0001
4	0.0018	0.0013	0.0001	0.0004
5	0.0013	0.0034	0.0016	0.0022

TABLE 6. **MAE test results (x10^2) across the different signature transactions and models.**

NASPERS LOB DATA					
Signature/ lagged order	XGB	RF	LR	DNN	AR
2	247.211	235.806	211.537	**98.603**	107.183
3	252.170	229.533	224.438	**94.441**	104.202
4	230.690	225.176	239.565	**100.480**	95.201
5	222.227	**219.592**	246.998	243.771	87.559

TABLE 7. **RMSE test results (x10^2) across the different signature transactions and models.**

NASPERS LOB DATA					
Signature/ lagged order	XGB	RF	LR	DNN	AR
2	285.108	256.440	225.921	**123.365**	129.952
3	293.676	251.296	242.385	**120.281**	127.077
4	271.017	246.373	261.568	**116.510**	127.379
5	**262.774**	273.798	311.198	372.300	107.048

TABLE 8. **MAE Test Results (x10^2) across the different signature transactions and models.**

AMAZON LOB DATA					
Signature/ lagged order	XGB	RF	LR	DNN	AR
2	14.896	**13.301**	13.395	17.032	14.159
3	14.586	13.785	**13.701**	17.607	14.164
4	14.319	**13.971**	14.722	15.120	14.191
5	13.893	**13.315**	14.153	15.644	14.168

TABLE 9. **RMSE Test Results (x10^2) across the different signature transactions and models.**

AMAZON LOB DATA					
Signature/ lagged order	XGB	RF	LR	DNN	AR
2	18.016	**15.956**	16.028	20.797	16.159
3	17.778	16.647	16.355	21.317	**16.201**
4	17.527	16.623	17.301	17.792	**16.237**
5	16.832	**16.185**	16.895	18.930	16.197

measurements. The difference in performance was due to a better model fit on AMAZON versus NASPERS where we find that RF performs better in developed markets while DNN performs better in emerging markets. Using the XGB model, we often see that the increase in truncation or order of signature results in a slight increase in performance in prediction errors. The signature approach represents a non-parametric way of extracting characteristic features from data; thus the higher the order of truncation, the more the information extracted by the signature, hence the better the expected performance by the models. Table 6 and Table 7 show a mixed performance between the machine learning models. DNN performs better at lower signature orders, while the RF and XGB model performs best at higher ones. In Table 7, DNN shows the best performance at most signature orders.

Table 8 shows a more consistent performance between the machine learning models. The linear regression model shows the best performance at signature order $n = 3$, while the random forest model shows the best performance at on other signature orders. This may be due to an increase in the number of coefficients when the order of signatures increases. Table 9 shows mixed results between RF and AR at different signature/lagged truncations.

To measure the performance enhancement of using signature features in machine learning algorithms, we tested the performance of Sig-DNN and Sig-RF, which are DNN and RF models with signature features, against DNN and RF models using raw LOB data. We used signature order 2, as it was the best performing signature order. Table 10 shows the prediction error results and Diebold-Mariano test p-values on Amazon LOB data, while Table 11 shows the results for NASPERS, and Table 12 shows the model efficiency in terms of average runtime for 30 iterations.

The p-values indicate statistical significance in the difference between the two sets of forecasts in both emerging and developed markets. Sig-DNN outperforms DNN in terms of accuracy and efficiency (less time). The model efficiency

TABLE 10. **The prediction error results (x10^2) and Diebold-Mariano test p-values on Amazon LOB data.**

AMAZON LOB DATA			
Model	MAE	RMSE	DM (p-value)
DNN	361.645	364.217	0.00002
Sig-DNN	**98.603**	**123.365**	
RF	**28.968**	**36.077**	0.00051
Sig-RF	235.806	256.440	

TABLE 11. **The prediction error results (x10^2) and Diebold-Mariano test p-values on NASPERS LOB data.**

NASPERS LOB DATA			
Model	MAE	RMSE	DM (p-value)
DNN	361.645	364.217	0.000023
Sig-DNN	**98.603**	**123.365**	
RF	**28.968**	**36.077**	0.000109
Sig-RF	235.806	256.440	

TABLE 12. **Model Efficiency performance (sec).**

Model Efficiency (Runtime)		
Model	NASPERS	AMAZON
DNN	18.696	18.960
Sig-DNN	**17.695**	**17.054**
RF	2.300	1.706
Sig-RF	**1.916**	**1.695**

of Sig-DNN is expected since LOB is compressed into few signature coefficients describing the path versus using raw LOB data. Sig-RF underperformed RF in accuracy but not efficiency, indicating that signature features do not work well with RF when seeking to reduce model prediction error in this setting. These results are consistent in all market types, i.e., developed and emerging markets. Overall, we find that a signature transform of LOB data for stock prediction improves DNN and not RF but still improves efficiency on both models.

VI. CONCLUSION

In this study, we have presented an approach to stock price prediction using signature path features of limit order book data. The results have been mixed using LOB data from two different exchanges at different times. While the earlier data from AMAZON performed better, the more recent NASPERS stream data showed that signature paths do not perform well as features for LOB stream data to stock price prediction due to higher model prediction error.

Our findings also reveal that the models tend to over-fit more with the AMAZON LOB data compared to the other LOB stream data used. This discovery suggests that careful selection of signature truncation or regularization techniques may be necessary to mitigate over-fitting and improve the out-of-sample performance of the models.

We also tested the potential enhancement of using signature path features versus raw LOB data. Our findings indicate a statistically significant difference between the forecasts generated by models using signature inputs and those using raw LOB data. The DNN model improved in terms of both accuracy and efficiency when using signature path features, while the RF model did not improve in terms of accuracy but did improve in terms of efficiency. The results demonstrate that using signature path features in DNN is better suited to capturing path information compared to raw data.

Given that we have established the use of signature techniques for feature generation and mean-mid price prediction, market makers can find optimal ways to hold inventory when trading. However, more tests are needed, and interesting questions remain that we hope to address in the future. For instance, can further increases in signature truncation or other LOB features prevent over-fitting of the various machine learning models? Similarly, are there possible investment strategies from the mean-mid price prediction, and are these strategies possible to implement practically given the high frequency nature of the data and the speed at which signature features are calculated at different truncations? Could the performance of the proposed framework be enhanced by obtaining data from the higher levels of the LOB?

Overall, our study shows that signature path features have potential in stock price prediction when used with machine learning algorithms, and their use should be explored further.

REFERENCES

[1] F. Abergel, M. Anane, A. Chakraborti, A. Jedidi, and I. M. Toke, *Limit Order Books*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[2] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison, "Limit order books," *Quant. Finance*, vol. 13, no. 11, pp. 1709–1742, Nov. 2013.

[3] L. G. Gyurkó, T. Lyons, M. Kontkowski, and J. Field, "Extracting information from the signature of a financial data stream," 2013, *arXiv:1307.7244*.

[4] T. Sidogi, R. Mbuvha, and T. Marwala, "Stock price prediction using sentiment analysis," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 46–51.

[5] T. Sidogi, W. T. Mongwe, R. Mbuvha, and T. Marwala, "Fusing sell-side analyst bidirectional forecasts using machine learning," *IEEE Access*, vol. 10, pp. 76966–76974, 2022.

[6] A. Ranaldo, "Order aggressiveness in limit order book markets," *J. Financial Markets*, vol. 7, no. 1, pp. 53–74, Jan. 2004.

[7] Y. Yu, "The limit order book information and the order submission strategy: A model explanation," in *Proc. Int. Conf. Service Syst. Service Manag.*, vol. 1, Oct. 2006, pp. 687–691.

[8] C. Cao, O. Hansch, and X. Wang, "The information content of an open limit-order book," *J. Futures Markets*, vol. 29, no. 1, pp. 16–41, Jan. 2009.

[9] T. Cenesizoglu, G. Dionne, and X. Zhou, "Effects of the limit order book on price dynamics," *Available at SSRN 2523643*, 2014.

[10] B. Zheng, E. Moulines, and F. Abergel, "Price jump prediction in limit order book," 2012, *arXiv:1204.1381*.

[11] J. Liu and S. Park, "Behind stock price movement: Supply and demand in market microstructure and market influence," *J. Trading*, vol. 10, no. 3, pp. 13–23, Jun. 2015.

[12] C. A. Parlour and D. J. Seppi, *Limit Order Markets: A Survey*. Amsterdam, The Netherlands: Elsevier, 2008.

[13] N. Popper. (Oct. 2012). *High-Speed Trading no Longer Hurtling Forward*. [Online]. Available: https://www.nytimes.com/2012/10/15/business/with-profits-dropping-high-speed-trading-cools-down.html

[14] R. S. Miller and G. Shorter, *High Frequency Trading: Overview of Recent Developments*, vol. 4. Washington, DC, USA: Congressional Research Service Washington, 2016.

[15] I. Aldridge, *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, vol. 604. Hoboken, NJ, USA: Wiley, 2013.

[16] J. Brogaard, T. Hendershott, and R. Riordan, "High-frequency trading and price discovery," *Rev. Financial Stud.*, vol. 27, no. 8, pp. 2267–2306, Aug. 2014.

[17] D. T. Tran, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Tensor representation in high-frequency financial data for price change prediction," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2017, pp. 1–7.

[18] A. Briola, J. Turiel, and T. Aste, "Deep learning modeling of limit order book: A comparative perspective," 2020, *arXiv:2007.07319*.

[19] A. Ntakaris, M. Magris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods," *J. Forecasting*, vol. 37, no. 8, pp. 852–866, Dec. 2018.

[20] P. Nousi, A. Tsantekidis, N. Passalis, A. Ntakaris, J. Kanniainen, A. Tefas, M. Gabbouj, and A. Iosifidis, "Machine learning for forecasting mid-price movements using limit order book data," *IEEE Access*, vol. 7, pp. 64722–64736, 2019.

[21] N. Passalis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Temporal bag-of-features learning for predicting mid price movements using high frequency limit order book data," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 4, no. 6, pp. 774–785, Dec. 2020.

[22] B. Hambly and T. Lyons, "Uniqueness for the signature of a path of bounded variation and the reduced path group," *Ann. Math.*, vol. 171, no. 1, pp. 109–167, Mar. 2010.

[23] D. Levin, T. Lyons, and H. Ni, "Learning from the past, predicting the statistics for the future, learning an evolving system," 2013, *arXiv:1309.0260*.

[24] T. Lyons, "Differential equations driven by rough signals," *Revista Matemática Iberoamericana*, vol. 14, no. 2, pp. 215–310, 1998.

[25] H. Buehler, B. Horvath, T. Lyons, I. Perez, and B. Wood, "Generating financial markets with signatures," Risk.net, Jun. 2021.

[26] W. Yang, L. Jin, H. Ni, and T. Lyons, "Rotation-free online handwritten character recognition using dyadic path signature features, hanging normalization, and deep neural network," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 4083–4088.

[27] J. Morrill, A. Kormilitzin, A. Nevado-Holgado, S. Swaminathan, S. Howison, and T. Lyons, "The signature-based model for early detection of sepsis from electronic health records in the intensive care unit," in *Proc. Comput. Cardiol. (CinC)*. IEEE, 2019, pp. 1–4.

[28] W. Yang, T. Lyons, H. Ni, C. Schmid, and L. Jin, "Developing the path signature methodology and its application to landmark-based human action recognition," in *Stochastic Analysis, Filtering, and Stochastic Optimization*. Berlin, Germany: Springer, 2022, pp. 431–464.

[29] P. Moore, T. Iliant, F. Ion, Y. Wu, and T. Lyons, "Path signatures for non-intrusive load monitoring," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3808–3812.

[30] T. Lyons, H. Ni, and H. Oberhauser, "A feature set for streams and an application to high-frequency financial tick data," in *Proc. Int. Conf. Big Data Sci. Comput.*, Aug. 2014, pp. 1–8.

[31] R. Cont, "Empirical properties of asset returns: Stylized facts and statistical issues," *Quant. Finance*, vol. 1, no. 2, pp. 223–236, Feb. 2001.

[32] T. J. Lyons, M. Caruana, and T. Levy, *Differential Equations Driven by Rough Paths*. Berlin, Germany: Springer, 2007.

[33] M. O'hara, *Market Microstructure Theory*. Hoboken, NJ, USA: Wiley, 1997.

[34] L. Harris, *Trading and Exchanges: Market Microstructure for Practitioners*. Oxford, U.K.: OUP, 2003.

[35] F. Diebold and R. Mariano, "Comparing predictive accuracy," *J. Bus. Econ. Statist.*, vol. 20, no. 1, pp. 134–144, Jan. 2002.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

THENDO SIDOGI was born in Venda, Limpopo, South Africa. He received the B.Sc. degree (Hons.) in mathematical techniques and the master's degree in mathematics from the University of Witwatersrand. He is currently pursuing the Ph.D. degree in machine learning in finance with the University of Johannesburg. He is also a Chartered Financial Analyst (CFA) Charterholder. His research interest includes the use of machine learning methods to improve quantitative investment methods.

WILSON TSAKANE MONGWE was born in Tembisa, Gauteng, South Africa. He received the B.Sc. degree (cum laude) in computing from the University of South Africa, the B.Bus.Sci. degree (Hons.) in actuarial science and in statistics and the master's degree (Hons.) in mathematical finance from the University of Cape Town, and the Ph.D. degree specializing in Bayesian machine learning from the University of Johannesburg. He was a recipient of the Google Ph.D. fellowship in machine learning, which supported his Ph.D. research. His research interests include computational finance, Bayesian machine learning, and Markov chain Monte Carlo algorithms.

RENDANI MBUVHA was born in Venda, Limpopo, South Africa. He received the B.Sc. degree (Hons.) in actuarial science and statistics from the University of Cape Town, and the M.Sc. degree in machine learning from the KTH Royal Institute of Technology, Sweden. He is currently the DeepMind Academic Fellow of Machine Learning with the Queen Mary University of London and an Associate Professor in statistics and actuarial science with the University of Witwatersrand. He is a fellow of the Institute and Faculty of Actuaries, U.K., and the Actuarial Society of South Africa, and a holder of the Chartered Enterprise Risk Actuary designation. He was a recipient of the Google Ph.D. fellowship in machine learning, which supported his Ph.D. research with the University of Johannesburg.

PETER OLUKANMI received the B.Sc. degree in systems engineering from the University of Lagos, the M.Sc. degree in computer science from the University of KwaZulu-Natal (UKZN), and the Ph.D. degree from the University of Johannesburg. His research interests include fundamental and applied data science and mathematical modeling. He is a winner of two IEEE conference awards in the field of soft computing and machine intelligence.

TSHILIDZI MARWALA (Senior Member, IEEE) was born in Duthuni, Venda, Limpopo, South Africa, in July 1971. He received the bachelor's degree in mechanical engineering from Case Western Reserve University, Cleveland, OH, USA, in 1995, the master's degree in mechanical engineering from the University of Pretoria, Pretoria, South Africa, in 1997, and the Ph.D. degree in engineering from the University of Cambridge (St Johns College), Cambridge, U.K., in 2000. He is currently a registered Professional Engineer (Pr. Eng.) with the Engineering Council of South Africa. He is also the rector of the United Nations University and an Under-Secretary-General of the United Nations. He has also completed other leadership courses from Columbia Business School, National University of Singapore, GIBS University of Pretoria, Harvard Business School, and University of South Africa. Some of his accomplishments include being a fellow of Cambridge Commonwealth Trust, in 1997, CSIR, in 2005, South African Academy of Engineering, in 2007, TWAS, The World Academy of Science, in 2010, African Academy of Science, in 2013, and South African Institute of Electrical Engineers, in 2016. His contributions in artificial intelligence field come in forms of over 15 books, and he has authored, over 50 peer-reviewed chapters, over 50 journal publications, and over 150 conference publications. He also holds the Deputy Chairman of the Presidential Fourth Industrial Revolution Commission of South Africa position and has contributed to over 50 articles to local press and journals on the subject of Fourth Industrial Revolution and Economics. He was the Vice-Chancellor and a Principal of the University of Johannesburg. He has won many awards, including notably, the Bronze Order of Mapungubwe awarded by the President of the Republic of South Africa, in 2004, and the TWAS-AAS-Microsoft Award for Young Scientists, in 2009.