Research in International Business and Finance 66 (2023) 102078

Contents lists available at ScienceDirect

Research in International Business and Finance

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

Full length article

Identification of high-frequency trading: A machine learning approach

Mostafa Goudarzi [a,b], Flavio Bazzana [a,*]

[a] Department of Economics and Management, University of Trento, Trento, Italy
[b] LUT Business School, LUT University, Lappeenranta, Finland

ARTICLE INFO

Keywords:
Market microstructure
High-frequency trading
FinTech

ABSTRACT

This study aims to develop a probabilistic model using machine learning techniques to identify high-frequency trading (HFT) based on order book data. The model enables precise intraday identifications, addressing the lack of a widely accepted framework for HFT identification and the inconsistencies arising from proxy indicators. Leveraging academic data, the model offers improved consistency and reproducibility for future HFT research. By incorporating fuzzy logic, the probabilistic model allows policymakers greater flexibility in shaping policies. The study utilises data from the BEDOFIH database of the French capital market and develops a robust classification model capable of accurately distinguishing HFT. Additionally, reverse engineering enhances the model's interpretability by transforming it into an interpretable regression tree without compromising its predictability. This research contributes to advancing HFT research, providing valuable insights, and offering a transferable methodology for identifying HFT in diverse market contexts.

1. Introduction

During the last few decades, there has been a keen interest in technological development, allowing industries to exploit their advancements to stay competitive in the highly competitive business environment. Finance, specifically financial markets, is one of the many industries that adopted information technology significantly, evolving into new electronic markets. The electronic platform eradicated all paperwork for buying and selling stocks and cut the trading process to a fraction of a second. This allowed market makers and other traders to increase the frequency of order submission within a relatively limited trading hour. As the trading frequency has increased, computers have been progressively used to send many orders to the market, providing a round-trip execution time of microseconds, laying the groundwork for the advent of High-Frequency Trading (HFT henceforward).

HFT is a breakthrough in finance that is constantly adopting innovations to enable ultra-fast data transmission because of a competitive nature against latency (Bernales, 2019). The (SEC, 2010) regarded High-Frequency Trading as one of the most significant market-structuring changes that accounted for more than 50% of US-listed equity trading. Yet, there is no unanimity among regulators or academics regarding the definition of high-frequency trading (Brogaard and Garriott, 2019). However, due to the increasing attention of scholars and regulators devoted to HFT, its operation has come under great scrutiny (Korajczyk and Murphy, 2019; Kelejian and Mukerji, 2016; Menkveld, 2016), especially concerning the impact on liquidity (Ammar et al., 2020; Yang et al., 2020).

The main characteristics of HFT identified by SEC's Concept Release (2010) are (i) the use of sophisticated and high-speed systems to place orders, (ii) investing in co-location facilities and dedicated data feed, (iii) initiating and concluding positions frequently

* Corresponding author.
E-mail addresses: mostafa.goudarzi@lut.fi (M. Goudarzi), flavio.bazzana@unitn.it (F. Bazzana).

https://doi.org/10.1016/j.ribaf.2023.102078
Received 30 November 2022; Received in revised form 21 July 2023; Accepted 27 August 2023
Available online 30 August 2023

and in very short timeframes, (iv) placing numerous orders and cancelling them in a fraction of a second and (v) maintaining zero or a low inventory at the end of the day.[1]

As a result of the lack of a standard definition for HFT, regulators and researchers have applied different identification methods. In general, regulators have greater access to data on trading venues, and their way of identification typically includes account-level information. For example, the European Securities and Markets Authority, in its report published in 2014, identified HFT in two steps: first, they flagged HFT companies based on their primary business, and then, based upon some criteria[2] on their order live, they specified the HFT activities. Indeed, they incorporate both direct and indirect methods for detecting HFT activities at the account level of each participant in the market (Benos and Sagade, 2016; Breckenfelder, 2019). On the other hand, academic studies conducted in the field of HFT have relied on various identification methods according to the quality and the attributes of the available data (Bazzana and Collini, 2020; O'Hara et al., 2019).

Some issues with these two identification methods constitute the primary focus of this research. Direct identification is carried out based on the information provided by HFT-based firms and excludes participants who do not reveal their involvement with HFT. For instance, HFT desks of investment banks with significant HFT activities are excluded in the direct identification (Biais et al., 2014). Those with privileged access to the account level data integrated some quantitative criteria with direct identification to identify HFT activities (Boehmer et al., 2018; Brogaard et al., 2014, 2017; Hagströmer et al., 2014; Hagströmer and Nordén, 2013; Kirilenko et al., 2017). A direct label of HFT has been applied in several studies with privileged access to venues' data at the account level (Hossain, 2022; ASIC, 2015; Breckenfelder, 2019; Brogaard and Garriott, 2019; Jarnecic and Snape, 2014).

In contrast, indirect identification tends to be more inclusive than direct identification because it is based on functional rather than institutional information. Different methods have been used in the literature to identify HFT based on trade data. Hasbrouck and Saar (2013) define a measure called *RunsInProcess* based on the practice of high-frequency traders of linking orders. Ersan and Ekinci (2016) adapt this measure to the order data from the Istambul market. Both Van Ness et al. (2015) and Aitken et al. (2018) build a proxy based on cancellation orders, Comerton-Forde et al. (2018) identify HFT using two measures of reaction speed. Brogaard and Garriott (2019) focus on the behaviour related to the overnight position, used with some adjustment also by Kang et al. (2022). Both Li (2021) and Ekinci and Ersan (2022) use the approach of inference from electronic message data, following Ekinci and Ersan (2018). Ekinci and Ersan (2018) introduce a measure called the HFT activity index (HAI), which captures the extent to which trading strategies rely on speed and technology. The authors test their approach using data from Borsa Istanbul and find that it effectively identifies HFT activities. Malinova and Park (2020) have identified HFT looking at orders by the same trader in different markets, i.e. the "sniping" strategy. More recently, Hossain (2022) used four different measures to verify their different power to identify HFT.

A different approach is based on the machine-learning framework. Mankad et al. (2013) propose a dynamic machine-learning method to uncover and analyse the ecosystem of an electronic financial market. It aims to identify and understand the relationships among various market participants, such as high-frequency traders, liquidity providers, and other market agents. Han et al. (2022) propose an explainable machine learning framework for discovering the dynamics of high-frequency trading in financial markets. The authors argue that traditional machine learning methods for HFT are often black-box models that need more transparency and interpretability.

All the studies that have used indirect methods to identify HFT have only used a single proxy or a small set of proxies to capture the operational characteristics of HFT (Hasbrouck and Saar, 2013; Leone and Kwabi, 2019; Mankad et al., 2013; Scholtus et al., 2014). Moreover, each study's proxies are usually different from those of others due to the dataset's unique features. The level of dispersion in proxy has contributed to inconsistency in empirical studies on HFT as Elizarov et al. (2017) have shown that identification of HFT is sensitive to the cutoff level of a particular proxy, let alone a different proxy.

This viewpoint guides our search for an identification strategy that meets two criteria. First, instead of relying on a small subset of proxies, it should initially make use of as much of the information provided in the order as possible. Second, it should make minimal use of account-level data or, at the very least, data that is easily accessible also to non-regulatory staff or academics. Additionally, we utilised fuzzy logic and immediate identification to outperform the regulator's identification. In contrast to binary logic, fuzzy logic allows for greater flexibility in HFT activity identification and policy application. As a result of this instant recognition, policymakers can restrict HFT activity in real time to exert more or less influence over the market.

In light of this perspective, the study employs machine learning to leverage a wealth of data for real-time identification. We used the label supplied by the French regulator to train the machine and elicit a general probabilistic model to address the identification inconsistency in prior research. The inconsistency stemmed from the need for a standard model for HFT identification, which caused scholars to use different proxies for HFT.

Since this study's probabilistic model is developed using data accessible to academics, it is expected to provide more consistency and reproducibility for further HFT research. The data for this research provide a classification based on functional and institutional information during a year. This classification is assumed as a target to be reached immediately following each order using our machine learning technique but with publicly available data. In other words, the machine learns from the classification made by the French regulator and produces a probabilistic model with publicly available features to classify any new order instantly. The final model incorporates all functional and institutional information at account level data into a trained classifier, allowing further research on HFT to use public order book data to identify HFT activities.

[1] Van Vliet (2017) introduces a simple model explaining the decision process made by HFT firms.

[2] When the 10th percentile of order life for a given firm in a particular stock is less than 100 microseconds, then that firm's trading activity in that specific stock is considered high-frequency trading (HFT).

Much literature is available on HFT, in which the emphasis is on the effect rather than the identification. In other words, most studies focus on HFT's influence on various market aspects rather than attempting to identify it in depth. Therefore, HFT has been extensively explored, but identification needs to be addressed. The purpose of this study is to fill this research gap.

In other words, this study contributes to the literature in two ways. First, it provides an instant and probabilistic method of identifying high-frequency trading on the French stock exchange using publicly available data. Second, and perhaps most importantly, our comprehensive coverage of the process, from feature engineering of row data to model interpretation, makes it straightforward to adapt and apply these techniques to identify HFT in any stock market.

2. Data

The data for this study comes from BEDOFIH.[3] This database includes historical high-frequency data from the main European stock markets, including Autorité des Marchés Financiers (AMF) Euronext Paris, which is opted for this research. The BEDOFIH AMF Euronext Paris database contains the order and trade histories of companies permitted to trade on Euronext Paris, whose market of reference is Euronext Paris.

The database contains all information about submitting, cancelling, modifying and executing every order in microsecond time stamps for each instrument per day. Each order has an ID which lasts until the complete execution or cancellation. It means the modification or partial execution of an order does not change the ID. Still, another variable (characteristic ID) keeps track of modification and increments by one in each step. This separation allows the analysis of each order modification during the time.

Furthermore, combining an order ID and the characteristic ID generates a unique key for the exploration. Although the database does not provide the data at the account level of market participants, it has used the information of each account to reveal the type of participants involved in each order and trade. The French regulator, the Autorité des Marchés Financiers (AMF), categorises all market participants into three groups: outright high-frequency traders (HFTs), mixed HFTs, and non-HFTs. More specifically, a market participant is recognised as an HFT if it has at least one of the following conditions (AMF, 2017): (i) the participant has cancelled at least 100,000 orders during the year, and the lifetime of those cancelled orders is below the average of the lifetime of all orders in the order book; (ii) the participant has cancelled at least 500,000 orders within 0.1 s after submission and 1% of those cancelled order persisted less than 0.0005 s.

The HFT distinguished by these conditions can be either an outright HFT or a mixed HFT. If a participant meets at least one of the conditions and is not an investment bank, it is labelled as outright HFT, but if the participant is an investment bank, it is categorised as mixed HFT. The third group of participants includes those who do not meet any of the two conditions and are flagged as Non-HFT.

Therefore, participants are classified based on information at the account level and in two steps sequentially. The first step undertakes the functional approach to differentiate all participants involved in high-frequency trading from non-HFTs. Then in the second step, an institutional approach is taken to classify those engaged in high-frequency trading to mixed and outright HFT based on their identity, whether it is an investment bank or not. Indeed, the identification of HFT activity is determined from a functional approach. The institutional approach in the second step is applied only to exclude the HFT desk of investment banks.

2.1. Sample

Rather than analysing individual transactions, this study examines orders submitted within the order book, regardless of whether they are executed, cancelled, or expired. Consequently, the sample encompasses all French-listed securities traded on May 4th, 2017.[4] This particular day yielded many order submissions, as indicated by the 867 order files in the BEDOFIH database.

Each order file within the dataset provides precise details about every order entered into the order book, including a timestamp accurate to the microsecond. Additionally, for each security traded at least once during the given day, a corresponding trade file contains supplementary information regarding the execution of orders, if applicable.

In the initial step, we exclude files that contain at most 150 orders or lack an accompanying trade file. This filtering process eliminates securities that had no trading activity or failed to attract a minimum of 150 orders. Consequently, our sample consists of 307 securities, collectively accounting for over 9 million orders.

Digging deeper into the sample's composition, it includes 12 bonds and 295 equities. Within the equity category, 97 equities fall into the micro-cap classification (with a market capitalisation below €250 million), 93 equities fall into the small-cap type (with a market capitalisation exceeding micro-cap but below €2 billion), 59 equities fall into the mid-cap classification (with a market capitalisation exceeding small-cap but below €10 billion), and 46 equities fall into the large-cap classification (with a market capitalisation exceeding €10 billion).

[3] European high-frequency financial database.

[4] We acknowledge the limitations inherent in our research, particularly about using data from 2017. It is essential to recognise that the strategies employed by high-frequency traders (HFT) may have undergone significant transformations since then, potentially influenced by the impacts of the pandemic on financial markets. However, it is crucial to emphasise that our research primarily aimed to identify individual orders submitted by HFT rather than analysing the overall patterns and dynamics of HFT activity. Regardless of the specific trading strategy, HFT traders employ the same means of executing their strategies: submission, modification, and cancellation of orders. Hence, our research focuses on analysing every single order independently from other orders, as these execution actions are the key components through which HFT strategies are implemented. We acknowledge that further research is necessary to delve into the strategic patterns of HFT activity and gain a more comprehensive understanding of its dynamics in contemporary market conditions.

Table 1
Features selected and extracted. The list of features with their codes and explanations. The first ten features are taken directly from the BEDOFIH database, and the others are extracted using our feature engineering approach.

Row	Feature code	Features explanation
1	o_cha_id	Characteristic identifier of the order starts at 1 and increments by 1 when the order is modified
2	o_state	How an order has left the order book
3	o_bs	Side of order: buy or sell order
4	o_type	Type of order: limit order, stop limit order or others
5	o_validity	Type of order validity: fill or kill, good for the day, etc.
6	o_price	Order price
7	o_q_ini	Initial order size
8	o_q_neg	Cumulative quantity negotiated for a given order
9	o_account	Type of user account (client account, own account, etc.)
10	o_nb_tr	Number of transactions attributed to the order
11	o_price_diff	Difference of order price with the last negotiated price
12	forecastError	Difference of order price with the next negotiated price
13	o_life	Life of order in millisecond
14	depth	Depth of order in the book at the time of submission
15	BBO_diff	Absolute difference of buy (sell) order price with then-current best bid (offer)
16	recs	If the order is submitted in trading group which is regulated covers equities with continuous trading and a static collar
17	o_member	Whether the order is placed with a high-frequency trader or not

Utilising the two trade and order files for the remaining securities, we constructed a dynamic order book by retiming and synchronising trades according to when the order was submitted. The resulting order book is at least as rich in information as the original order book. Unlike most other studies, our data are not adjusted to be distributed at equal intervals. This allowed us to preserve all of the information inherent in the dataset.

The features extracted from the constructed order book at the time of order submission are then-current depth, best bid/offer and the difference of price order with the prevailing best bid/offer and the last traded price. These features, along with some other available features at the time of submission, such as the number of times the order is previously modified, buy or sell order, type of order, order validity, order price, initial order size, and even account type cannot be differentiating in the kind of trader. It should be noted that several additional details revealed after the conclusion of the order (being cancelled, filled, or modified) enhance the data.

Our sample's median order life is 6.7 s and around one-third of orders last less than one second. It indicates few seconds after order submission, most orders leave the order book and reveal more information such as order life, type of book release (modified, fully filled or partially filled), number of transactions involved, number of modifications, a difference of order price with next negotiated price (as forecast error) and cumulative quantity negotiated. This indicates that it would be worthwhile to delay identification until the end of the life of the order, thus having richer information for more accurate classification, as the delay is considerably less than the existing annual identification.

As mentioned, the BEDOFIH database provides the classification for each order based on account-level information about the traders and annual investigation of their trade activities. This classification is performed once a year and requires traders' ID, which is not exposed to the public.

Orders without price, like market orders and orders submitted during the pre–opening, are excluded from the sample. Since most features are extracted based on the order price, marker orders submitted without expressing the price are excluded from the sample. Furthermore, charges submitted during the pre–opening period are excluded due to studies (Bellia, 2018) showing that traders perform differently due to the different nature of trading mechanisms that are not continuous auctions but double auctions.

2.2. Features engineering

Unlike transaction-based data, an order-based database contains many observations since many orders do not end up with transactions for many reasons. While each observation includes some information that cannot be removed, it is possible to draw some variables irrelevant to the study, increasing the computation's efficiency and preventing overfitting.

On the other hand, it is possible to create new variables from existing variables following the theory that provide more information than their initial counterparts. Therefore, we organised the database to include the following features using a process of elimination and creation of variables, known as feature engineering.

Table 1 reports the features selected for this study. Features 1 through 10 are taken directly from the database, while features 11 through 16 are extracted from other variables using feature engineering and reconstructing the order book. Features 11 and 12 reflect the difference between the order price and the last and next price at which the security is traded. The two features arise from the fact that those who immediately follow the market (with the lowest `o_price_diff`) differ from those who lead (with the lowest `forecastError`) the market. The order Life measures the interval between when an order is validated and when it leaves the order book. Historically, high-frequency traders have invested heavily to reduce their latency, which is the rationale behind this

feature. Due to the ability of HFTrs to monitor all bids and offers in a millisecond using powerful computers, we assumed `Depth` and `BBO_diff` could distinguish HFT from other traders. `recs`, the last feature extracted, indicates trading groups regulated and covering equity with continuous trading and a static collar. Along with the fact that HFTrs are interested in high-liquidity stocks due to the low-latency nature of their trading, we suggested that `recs` is the characteristic that differentiates trading groups with the best liquidity.

The idea of adding `recs` to the database comes from the fact that the initial investigation of this classification revealed that the participation of HFTs is very diverse across securities, with a min of 0 and a max of 99.77 per cent. The minimum (maximum) participation is in securities with a low (high) number of orders. For example, 287 orders have been submitted for FR0000188799 (an OAT Bond), all by non-HFTs, while 99.77 per cent of 95,919 orders submitted for FR0010221234 (Eutelsat share) belong to HFTs. There is no doubt that HFT plays an imperative role in causing high numbers of orders for a given security and may result in concerns regarding reverse causality. Still, to begin with, it is the features of security that attract HFTrs.

Euronext Paris has already considered these differences and placed every security in mutually exclusive trading groups. Every trading group has regulations and restrictions, some of which may decrease liquidity and discourage HFTers from participating. Since the trading groups in Euronext Paris are numerous, we created `recs` indicator to separate more liquid ones from others.

Considering that there are many trading groups at Euronext Paris, we created the `recs` indicator to distinguish the trading groups with higher liquidity from others. `recs` represents any regulated trading group in which Equities are traded with continuous auctions and static collars.

2.3. Order book reconstruction

The order book must first be reconstructed to extract features such as depth and `BBO_diff`. Consequently, each limited buy (sell) order is compared to all other buy (sell) orders already submitted and valid in the order book to find its rank inside the bid (offer) side of the order book. The task was quite demanding because the number of valid orders in the order book grew during the trading day, and each of the last orders must be checked against millions of valid orders. Comparisons are based on the priority of prices on two sides of the order book. We first developed an indicator for each order to identify which orders needed to be included in the comparison based on their type, time of submission, and validity. Second, we extracted and sorted the prices among the included orders and determined the ranking of the submitted order within them. The rank indicates the depth of the order at the time of submission.

As a result of having a list of flagged orders for each timestamp in which an order was submitted, we could determine the appropriate bid and offer for each submission time. Consequently, the `BBO_diff` was created, which indicates the difference between each order's price and the then-current best bid or offer. This feature is slightly different from `o_price_diff`, which indicates the price difference from the last traded price since the best bid or best offer may not result in a transaction.

Our study aims to include as much data as possible to train the machine without overfitting, so we examine all securities simultaneously. Despite this, the features of securities differ significantly in value. This issue has been addressed by standardising the features of securities so that they are placed into the same range of variability. For example, we scaled the order price, the initial quantity of an order, negotiated quantity and best bid and offer to their median.[5] While `o_price_diff` and `forecastError` scaled to the median of order price divided by 100 to be shown as a percentage of the price.

Another issue with our order data is the presence of extreme outliers. Despite their very high values, sometimes hundreds of times more than the median, they were comparatively small in number, so a 99.98% winsorization[6] was able to smooth out our data.

According to the BEDOFIH classification, mixed HFTs are investment banks that carry out HFT activities but are categorised in the same way pure HFTs are classified. Therefore, from a functional perspective, mix and pure HFT are equivalent. Consequently, we disregard this distinction and consider all pure and Mixed HFT to be part of one HFT class regardless of whether investment banks are involved.

2.4. Hidden orders

Various exchanges and trading venues offer the option to hide orders; however, it is essential to distinguish that complete order hiding is typically permitted in dark pools rather than in the lit markets, which are the primary focus of our investigation. In lit markets, there is usually a requirement to display a minimum quantity for orders, commonly known as iceberg orders.

Our dataset encompasses both disclosed and hidden quantities of iceberg orders. Iceberg orders, appropriately named, consist of a visible portion that appears on the order book, similar to any other order. The visible portion of an iceberg order is placed in the order book based on its price, featuring a lower quantity than the overall order size. It is essential to highlight that the presence of hidden portions within iceberg orders does not compromise the accuracy or validity of the features we extracted from the order book, which depends on the price of the order. For instance, the depth of each order in the order book is influenced by its price and the prices of other orders regardless of the number of orders.

[5] Since most orders have zero negotiated value, we ignore zero when calculating the median for negotiated quantity.

[6] For values less than percentile 0.01, the percentile was replaced, and for values greater than 99.99%, the corresponding percentile was replaced.

Table 2
Summary statistics of numerical variables grouped in HFT and Non-HFT. N is the number of observations. For each feature, there are two rows. The first row shows the summary statistics of the HFT group, and the second is for the non-HFT group.

Variable	N	Median	Min	Max	1st quart.	3rd quart.	Mean	Sd
o_cha_id	8.77×10^6	1	1	6 961	1	1	7.96	119.3
	2.53×10^5	109	1	16 558	1	4 726	2 862	4 400
o_price_buy	4.40×10^6	0.999	0.689	1.225	0.995	1.002	0.998	0.009
	1.29×10^5	0.998	0.689	1.451	0.994	0.999	0.993	0.027
o_price_sell	4.37×10^6	1.001	0.698	1.451	0.997	1.004	1.001	0.009
	1.24×10^5	1.002	0.689	1.451	1.001	1.007	1.010	0.042
o_q_ini	8.77×10^6	1	0.004	303.6	0.699	1.48	1.70	5.36
	2.53×10^5	1.132	0.004	303.6	0.943	3.77	5.67	22.53
o_q_neg	8.77×10^6	0	0	127	0	0	0.15	1.56
	2.53×10^5	0	0	127	0	0	0.68	4.71
o_nb_tr	8.77×10^6	0	0	292	0	0	0.13	0.65
	2.53×10^5	0	0	402	0	0	0.39	3.98
o_price_diff	8.77×10^6	0	−1 660	1 550	−1.5	1.5	−0.39	42.96
	2.53×10^5	1	−1 660	1 550	−4	35	2.99	141
forecastError	8.77×10^6	0	−2 200	1 390	−2	1.6	−0.57	44.96
	2.53×10^5	0	−2 200	1 390	−8	23	0.40	163.9
o_life	8.77×10^6	6 223	0.001	1.95×10^{10}	265	42 110	6.36×10^5	3.30×10^7
	2.53×10^5	4 863 101	0.002	2.07×10^{10}	175 986	1.34×10^7	3.78×10^7	3.63×10^8
depth	8.77×10^6	2	1	173	1	4	4.88	9.81
	2.53×10^5	2	1	162	1	3	6.08	12.77
BBO_diff	8.77×10^6	2.2×10^{-4}	0	0.47	1.02×10^{-4}	6.34×10^{-4}	1.4×10^{-3}	6.7×10^{-3}
	2.53×10^5	1.1×10^{-3}	0	0.51	0	4.3×10^{-3}	0.011	0.036

To comprehensively address any concerns related to iceberg orders, we conducted a specific identification and analysis of these orders in our study. Our examination revealed that iceberg orders represent a small fraction, comprising approximately 1% of the total orders submitted in the limit order book. Additionally, we calculated the ratio of the hidden portion within iceberg orders within our sample. This ratio represents the percentage of the iceberg order's quantity that remains concealed and is not visible in the order book data. Our analysis demonstrated that the hidden ratio ranges from a minimum of 0.2% to a maximum of 99.96%, with an average value of 78%. Given that the hidden ratio does not reach 100%, we can confidently assert that all hidden orders in our dataset are partially hidden and none completely concealed.

Therefore, it is evident that the partially hidden orders in the lit markets uphold the effectiveness of the methods we have developed in this study for feature extraction and identification. Moreover, achieving real-time identification of HFT by considering the visible portion of the iceberg orders remains feasible.

3. Descriptive statistics

Table 2 shows the summary statistics of numerical variables grouped by HFT and non-HFT. For each variable listed in the left column, there are two rows such that the first row summarises the HFT orders, and the second reports the non-HFT orders. To see if there is a significant difference between the means of the two groups, we ran a two-sample t-test for all numerical variables. The null hypothesis was rejected at a very low (almost zero) significance level.

Table 3 reports the relative frequency of categorical variables in two groups of HFT and non-HFT orders. The values are reported as percentages and add up to 100 in each row. The relative frequency of `o_state` shows that the broker cancels most HFT orders while most non-HFT orders are modified. This is further supported by the higher mean and median of `o_cha_id` among non-HFT orders since the characteristic ID increments whenever the order is modified. `recs`, as an extracted indicator for securities with high liquidity, presents a great power of differentiating so that 99.59% of HFT activities took place in securities with `recs` indicators.

Tables 2 and 3 demonstrate that the selected and extracted features differ considerably between HFT and non-HFT orders. Thus, there is an opportunity to identify HFT by these features.

4. Method

Focusing on identifying HFT with the selected and extracted features, we try to find an algorithm of machine learning that is more suitable for the order data. The target we will reach by the ML algorithm is the AMF classification but with fuzzy logic. After finding the best algorithm for identification, we test the algorithm against other sets of data from other trading days. Ultimately, we interpret what the model has done in the identification process.

4.1. Choosing an algorithm

The main objective of this study is to develop a probabilistic model to identify high-frequency trading with no need for trader ID at the earliest possible time. Current identification models require trader ID, which is only available to the exchange authorities and is performed once a year since they are obtained by annual data. In particular, we aim to exploit the yearly and account-based

Table 3
Relative frequency of categorical variables categorised by HFT and Non-HFT members. Each row represents HFT or Non-HFT group, while the columns correspond to different categories. The values in each cell indicate the proportionate frequency (in per cent) of that particular category within the HFT or Non-HFT group. For example, in the top-left cell, it can be observed that 7.51% of HFT orders have a state of 2. The codes 2, 4, 5, C, 0, 3, S, and P in the `o_state` variable represent the following order states: totally filled, cancelled by the broker, modified, cancelled by the trading system, new entry in the book, eliminated due to day validity, eliminated by supervision, and cancelled by the self-trade prevention, respectively. In addition, the `o_validity` variable includes codes 0, 1, 2, 3, 4, 6, 7 representing good for the day, good till cancel, valid for auction, fill or kill, good until a specific date, and valid for closing, respectively. Similarly, the `o_account` variable uses codes 1, 2, 3, 4, 6, and 7 to indicate order submissions by client account, own account, retail liquidity provider, retail market organisation, liquidity provider, and parent company account. Moreover, the `o_type` variable employs codes 2, K, and 4 to denote limit orders, market-to-limit orders, and stop limit orders, respectively. The `o_bs` variable uses B and S to indicate sell and buy orders, respectively. Lastly, `recs` refers to orders submitted for an asset in a regulated trading group that covers equity with continuous trading and a static collar.

`o_state`	2	4	5	C	0	3	S	P
HFT	7.51	78.33	13.91	0.14	0.01	0.06	0.00	0.03
Non HFT	16.12	12.50	66.46	2.35	0.41	2.16	0.01	0.00
`o_validity`	0	1	2	3	4	6	7	
HFT	96.67	0.00	0.00	3.08	0.14	0.10	0.00	
Non HFT	83.97	1.47	0.00	4.03	8.12	2.42	0.00	
`o_account`	1	2	3	4	6	7		
HFT	4.84	25.48	8.19	0	61.36	0.13		
Non HFT	19.93	17.04	0	9.20	53.83	0.01		
`o_type`	2	K	4					
HFT	99.99	0.01	0.00					
Non HFT	97.80	1.44	0.77					
`o_bs`	B	S						
HFT	50.14	49.86						
Non HFT	51.06	48.94						
`recs`	false	true						
HFT	0.41	99.59						
Non HFT	54.70	45.30						

identifications that exchange authorities perform to develop a robust model capable of distinguishing HFT but only with publicly available data on order books.

As mentioned, we use order-level data in Euronext Paris provided with HFT flags. The data lacks trader ID for orders, and our analysis is based on intraday aggregated data in two groups of traders, high-frequency traders (HFTrs) and others called non-HFTrs. Although the French market regulator categorised HFT into two groups, Pure and Mix, depending on whether the trader is an investment bank, we ignored this further categorisation because Mix and pure HFT are classified by the same characteristics and expected to have the same functions.

The starting point for choosing a classification algorithm stems from the data characteristics that, in our case, are low dimensional (the number of observations is staggeringly high while few features are available) and highly skewed class distribution. Additionally, the interaction between the features is determinative and must be considered. These characteristics and requirements fit well in the decision tree approach.

Decision tree

A decision tree is a supervised learning algorithm of machine learning in which the machine splits the data into two or more in each step based on single features to improve the classification. Despite the simple logic, it has many advantages over other classification methods.

It makes no assumptions about relationships between variables, so there is no concern about collinearity which is likely in our data. For example, if two features are highly correlated and we split based on one, little or no information can be obtained by splitting on the other, and it will be ignored in favour of another feature.

The decision tree considers the interaction between features due to its sequential algorithms. The interaction is theoretically and practically significant because we have few features, and each interaction between a few variables may explain the variability in the classifications. Immunity of the decision tree to outliers is also crucial since the order book contains fleeting orders submitted by fat finger error or for manipulation. Unlike the regression model, the decision tree considers non-linearity and does not require scaling or converting categorical variables to several dummy variables.

Despite many advantages, it bears some disadvantages, among which overfitting is more critical. A single decision tree deeply grown with many branches has a shallow bias, but it might fail to predict out-of-training sample data. Several techniques have been suggested to overcome the vulnerability to overfitting, each of which mitigates one of the abovementioned advantages. We can compromise one characteristic in favour of more generalisability depending on which characteristic is more important in each situation. For example, when accuracy is more critical than interpretably, we implement the ensemble tree method, which addresses the skewness issues and will be discussed in the next section. Limiting depth, leaf size, and pruning are other techniques for overcoming techniques that improve generalisability and interpretability with a potential reduction in accuracy.

Ensemble tree

The ensemble is a supervised learning approach in which multiple learning models (weak learners) are combined to produce a better predictive performance. Supervised learning algorithms generally look for a hypothesis through a hypothesis space that best fits the data concerning the particular problem. Ensembles combine each algorithm's found hypotheses to form a better-performing hypothesis. This hypothesis may not be included in the hypothesis space of weak learners.

There are several types of ensembles, and choosing the best option depends on why we move from base learners to ensemble learners. In our case, the concern is the imbalanced number of classes and overfitting. Therefore, boosting is the best (so far) ensemble technique that takes care of these issues. It is noteworthy that ensemble methods typically require more computational and storage resources than base learners, especially regarding the order book data, which becomes more crucial. For example, among all six available boosting techniques, our resources could not afford to apply "bootstrap aggregating" (aka bagging or random forest) and "total boost" on our extensive dataset. However, they could have performed better in the reduced data version over the other four methods.

Therefore, we evaluate four boosting methods- adaptive boosting, logic boosting, robust boosting, and random under-sampling (RUS) boosting- all available in the MATLAB machine learning toolbox. The selected ensemble methods work sequentially so that the misclassification of each step will be used to adjust the hypothesis space of the next step. Each boosting method performs this sequential process differently.

Boosting evaluation

Theoretically, ensemble tree-boosting methods are evaluated by investigating the process of combining weak learners to see how they deal with different data characteristics such as noise, skewness, etc. Another evaluation method is to use real-world data and seek the best predictive power. Although the empirical approach requires more computational resources, the results are more realistic because no assumptions are made.

Following empirical evaluation, we employ the hold-out technique for validation, which is a part of the data excluded from the training to be used later to evaluate the trained model's precision. As already explained, data are cleaned from outliers and irrelevant features and another side, some features like order life and order depth were extracted and added to the database.

We created a non-stratified random partition to evaluate boosting methods that held 20% of data out of training. Holdout data will later be used as test data to assess and compare each model's predictive power. The standard CART algorithm constitutes 100 weak learners for an ensemble with surrogate splits. Due to missing values in the data, ensembles of trees with surrogate splits perform better.

The classification provided by BEDOFIH is multi-class, including pure-HFT, Mix-HFT and Non-HFT. Specifically, BEDOFIH classifies traders in two sequential steps. First, it functionally separates all HFTs from Non-HFT and then sub-classifies those HFTs into Mix-HFT and Pure-HFT based on whether they are investment banks. Since the investment bank affiliation is not available in order book data, we only focus on the first step, functional classification and differentiate Mix and Pure HFT from Non-HFT.

To Evaluate the classifier, one can use many measures, each driven from a different proportion of cells in the confusion matrix. Contrary to most classification problems focusing on precision measures, we pay attention to sensitivity and specificity measures in this stage because the goal is to find the best model with higher power of detectability rather than interpreting the predicted class.

Precision (aka positive predictive value), recall (aka sensitivity) and overall misclassification error are provided in each matrix. In the case of HFT class, precision answers "What percentage of predicted HFT are actual HFT?" and recall is the answer "What percentage of actual HFT are predicted as HFT?". While the nominator of both accuracy measures is the number of indeed predicted HFTs, the denominator of precision is the number of predicted HFTs, and recall is the number of actual HFTs.

Sensitivity, also known as recall, power of the test and True Positive Rate (TPR), shows what percentage of HFTs are detected as HFTs. On the contrary, specificity, also known as True Negative Rate (TNR), indicates the ratio of Non-HFT, which the model genuinely detects. Error type 1 is the complement of TNR, and error type 2 is the complement of TPR that is $\mathrm{TPR} = 1 - \beta$ and $\mathrm{TNR} = 1 - \alpha$.

Results of validation

We trained four models using data from May 4th, 2017 and 4 boosting methods for the ensemble tree. The validation of these models has been performed using hold-out data from the same day. Fig. 1 reports the validation of four trained models depicted in the confusion matrices. The matrices show how many order in the hold-out data is classified or misclassified as HFT and non-HFT. The numbers in these 2×2 matrices allow us to compute any evaluation metrics.

Model selection approach

As confusion matrices in Fig. 1 show, none of the models outperforms others in all validation measures. Thus, the approach for determining the best model is decisive in highlighting the critical measures.

As mentioned, in the model selection stage, the focus must be on the TNR and TPR, but even selecting between these two for the final decision on the best model requires a different approach. If the cost of misclassifying HFT (i.e. The true class is HFT but detected as NON) is high, TPR is the decisive measure. TPR shows what percentage of HFT is indeed detected by the trained model. Conversely, if the cost of misclassifying NON is high, TNR is the measure for finding the best model. This study's approach is on the high cost of misclassifying NON, which implies we prefer a model with the highest detection rate of NON without vital compromise in HFT detection.

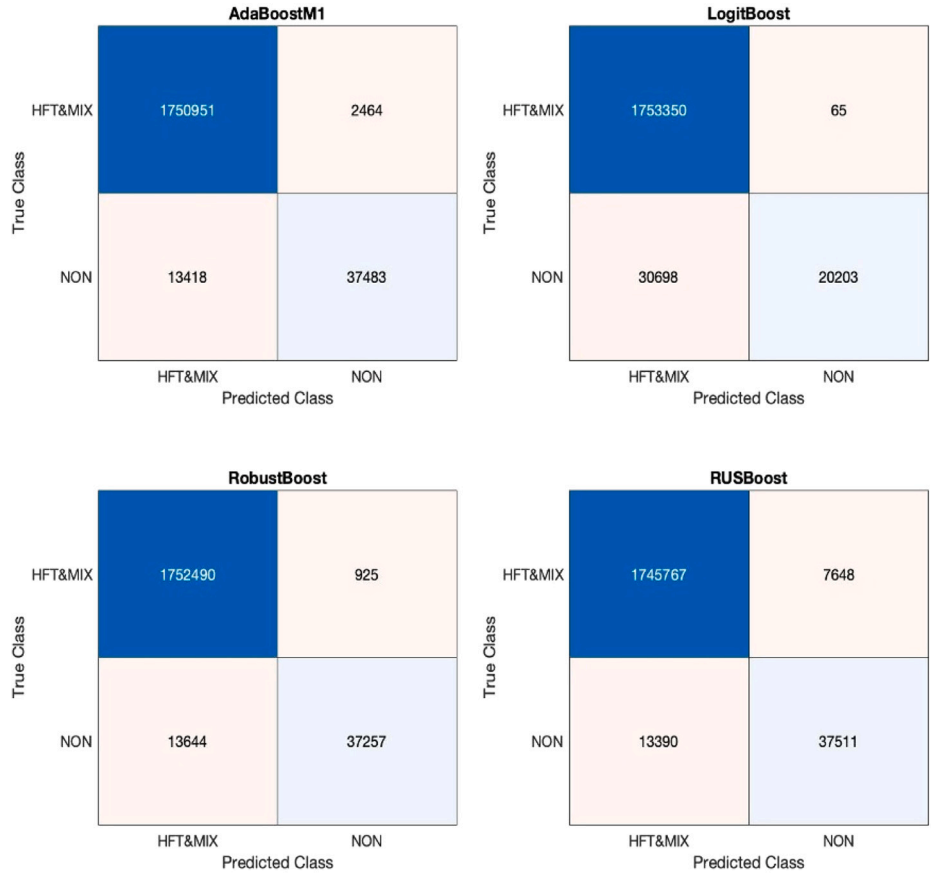

Fig. 1. Confusion matrices. It shows the model's confusion matrices trained with the ensemble tree's four-boost technique. Target classes are true member types of holdout data, and output is predicted member types by each model. The confusion matrices report the number of predicted versus actual classes. The true predictions are depicted in blue, and the wrong predictions are shown in pink. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 4
True negative and positive rate of four models against different sets of data.

		AdaBoostM1	LogitBoost	RobustBoost	RUSBoost
Against hould-out subset	TNR	0.73639	0.39691	0.73195	0.73694
	TPR	0.99859	0.99996	0.99947	0.99564
Against data from 02/01/2017	TNR	0.53894	0.08962	0.53794	0.63961
	TPR	0.99576	0.99981	0.99752	0.98808
Against data from 03/01/2017	TNR	0.56524	0.15148	0.55740	0.65282
	TPR	0.99693	0.99945	0.99786	0.99276
Against data from 04/01/2017	TNR	0.60478	0.17817	0.60202	0.68572
	TPR	0.99475	0.99855	0.99555	0.99024
Against data from 05/01/2017	TNR	0.63293	0.23851	0.63753	0.72329
	TPR	0.99605	0.99974	0.99706	0.99176
Against data from 06/01/2017	TNR	0.67545	0.38210	0.64299	0.72286
	TPR	0.99797	0.99994	0.99886	0.99418

Results of testing by out-of-sample data

Table 4 lists the TPR and TNR of the models. The first two rows are from the classifying test of 20 per cent of data that has been held out of training, and the other rows come from a test conducted on data from different days.

Throughout all tests, the model trained using the RUS boost ensemble technique provides the most significant rate of TNR while providing almost the same rate of TPR as other models. On average, choosing RUS boost over other models improves TNR by at least 8.14%, compromising only 0.45% of TPR. It is important to note that the higher TPRs can be attributed to the fact that HFTrs generally share similar trading activities and exhibit common characteristics. This homogeneity in HFT behaviour makes it easier for the model to identify and classify HFT instances accurately. On the other hand, the non-HFT category encompasses a diverse

Table 5
Predictive and Negative Predictive Values.

PPV and NPV	20% Hold-out	2.1.2017	3.1.2017	4.1.2017	5.1.2017	6.1.2017
PPV of RUS boost model	0.9924	0.9812	0.9847	0.9845	0.9853	0.9882
NPV of RUS boost model	0.8306	0.7385	0.8007	0.7764	0.8245	0.8418

Table 6
True negative and positive rate of four models against six sets of data.

		AdaBoostM1	LogitBoost	RobustBoost	RUSBoost
Against hould-out subset	TNR	0.70771	0.75197	0.75301	0.89767
	TPR	0.99252	0.99349	0.99267	0.95654
Against data from 04/05/2017	TNR	0.75276	0.76638	0.76947	0.87106
	TPR	0.99498	0.99457	0.99538	0.94946
Against data from 03/01/2017	TNR	0.65148	0.67438	0.68902	0.90164
	TPR	0.99326	0.99353	0.99357	0.95706
Against data from 04/01/2017	TNR	0.6845	0.70093	0.71558	0.90985
	TPR	0.99149	0.9919	0.99097	0.95125
Against data from 05/01/2017	TNR	0.70965	0.73087	0.74474	0.92327
	TPR	0.99251	0.99143	0.99278	0.95401
Against data from 06/01/2017	TNR	0.69898	0.7234	0.73186	0.89897
	TPR	0.99502	0.99525	0.99502	0.9557

range of market participants, including different trading strategies and objectives. This heterogeneity within the non-HFT group presents a more significant challenge for the model to detect non-HFT traders, leading to comparatively lower TNRs accurately.

The observed increasing trend in TNR from January 2nd to January 6th can be attributed to the higher homogeneity of order characteristics among non-HFT participants, particularly in a direction opposite to that of HFT traders. To provide a more straightforward explanation, let us consider a scenario where our model detects HFT traders solely based on the short lifespan of their orders. A higher TNR on a specific trading day indicates that the lifespan of orders submitted by non-HFT participants is longer and more uniformly distributed among all non-HFT traders. In other words, there is a greater homogeneity among non-HFT participants in terms of order duration, specifically in a manner that distinguishes them from HFT traders. As a result, increasing the homogeneity among non-HFT participants in the opposite direction of HFT detection leads to a higher TNR.

Given these observations, we have chosen the RUS boost[7] for the rest of our analysis. Table 5 shows the positive and negative predictive values (PPV and NPV) for the RUS boost model against hold-out data and five other data sets from different days. PPV and NPV describe the model's predictive performance in detecting HFT and Non-HFT. For example, a PPV of 99.24% explains that less than one per cent of those seen as HFT is misclassified. The NPV is less than PPV, meaning non-HFT orders cannot be detected accurately by those features we chose and extracted for HFT identification. It can be because non-HFTrs are more diverse than HFTrs regarding trading strategies. Nevertheless, the model exhibits satisfactory performance in identifying non-HFT instances.

The results highlight the selection of the boosting method based on the high penalty associated with misclassifying non-HFTrs, while still providing a robust ability to identify HFTrs.

Therefore, we suggest an ensemble decision tree boosted by the RUS method for identifying HFT. Although the model is highly accurate, little information can be gleaned from it to determine how the identification has been carried out and how it relates to the theoretical definitions of HFT. To address this issue, we will present some methods for interpreting the final model in the next section.

Robustness check

We reproduced everything we had done up to this point, but this time we trained the models using data from January 2, 2017, to see if our results would be robust if we trained the models using a different data set.

Table 6 demonstrates the consistent superiority of the RUS boost model over the other three models in terms of true negative rate, even when trained with different datasets. This performance advantage comes with only a minimal compromise in true positive rate. Consequently, our findings exhibit robustness across diverse datasets used for training and evaluation.

The persistent higher true positive rate compared to the true negative rate can be attributed to the homogeneity of HFTrs, as we have previously discussed. In contrast, the heterogeneity of non-HFT participants contributes to the relatively lower true negative rate. This phenomenon aligns with our earlier explanation of the characteristics and behaviours exhibited by HFT and non-HFT traders.

Overall, these results reinforce the strength and generalisability of our findings, indicating that the RUS boost model consistently outperforms alternative models in capturing true negatives while maintaining competitive true positive rates.

[7] The problem of random undersampling adaptive boosting (RUS boost) is that it only considers a small part of data for training. For example, if the class with fewer members consists of 3 per cent of data, it undersamples the more significant class so that two classes have the same number in training. Due to our numerous observations, we could train our model correctly even though it covered only a tiny fraction of our dataset.

Table 7
Performance of RUS-boost model trained with different datasets for different market capitalisation.

	Subset proportion	Trained by data from May 4th 2017		Trained by data from Jan 2nd 2017	
		F-score	Accuracy	F-score	Accuracy
Against Large-cap subset	70%	0.9954	0.9909	0.9774	0.9561
Against Mid-cap subset	25%	0.9952	0.9904	0.9749	0.9514
Against Small-cap subset	4%	0.9770	0.9564	0.9170	0.8537
Against Micro-cap subset	1%	0.8923	0.8476	0.7730	0.7257

To assess the performance of our RUS-boost model in different market capitalisation categories, we trained the model using two distinct datasets. Table 7 provides an overview of the model's performance across various subsets and includes the proportion of orders in the order book for each category.

The results showcase the model's capability to identify HFT orders across different market segments effectively. With significantly high F-scores and accuracy, the model performs exceptionally well for the large-cap and mid-cap subsets, which constitute the majority of orders in the order book (70% and 25%, respectively).

While the model's performance remains commendable for the small-cap subset, there is a noticeable decrease in performance as the subset proportion decreases. This trend becomes more pronounced in the micro-cap subset, representing a mere 1% of orders, where the model faces a more significant challenge. The lower F-score and accuracy values in this subset can be attributed to the limited data available for micro-cap orders. The smaller proportion of orders poses a difficulty for the model in accurately identifying HFT activity within this category.

Overall, the analysis affirms the effectiveness of the RUS-boost model in identifying HFT orders across diverse market capitalisation categories. It emphasises the importance of considering the subset proportions and data availability when interpreting the model's performance.

4.2. Interpretation of the model

As already mentioned, ensemble decision tree classifiers are generally more accurate than simple trees, but they lead to predictors that are difficult to interpret compared to decision trees. Therefore, it is compelling to understand how the classifiers made their classifications. One way to achieve this is to reverse engineer the process to move as far backwards as possible.

Reverse engineering

Accuracy and interpretability are two aspects that hardly come together, so we take a reverse engineering approach to explain a highly accurate model.

The ensemble tree returned a model with high accuracy but uninterpretable. An interpretable model with clear logic can illuminate the ensemble model's black box and clarify clear reasoning. The starting point for this reverse engineering process is those estimated scores from the ensemble model. The model produces a matrix of scores with one row for each observation and two columns. Although the mathematical analysis of how these scores are calculated is beyond the scope of this paper, we know the score of each observation represents the degree of confidence that the observation is part of that class. The higher the score, the greater the level of confidence.

These scores are, in fact, one step before a final binary classification is made. Based on an observation's features, the ensemble model assigns scores to each observation and then classifies the observation by the resulting score. The investigation is performed in a convoluted sequence of processes called a black box. Various ensemble methods generate scores based on the other methods used to calculate them. In the case of RUS boost, the scores are all positive, and the sum of each observation's scores for two classes is always a constant. If the observation's score for a given class is higher than the average of its two scores, the model classifies the observation in that given class otherwise in the other class. The magnitude of the scores indicates the degree of confidence in the classification. Scores around the average are those observations classified with the lowest confidence.

The threshold simplifies classifying observations but keeps us from knowing how confident we are about each observation class. This means that in the transition from scores to binary classification, we lose the confidence levels for each class. Indeed, scores are richer in information than the final binary classification. In the next section, we will use these scores to create a simulated dataset to understand the black box better.

Simulation

To determine how the calculated scores come from the features, we create a dataset with the features and the score based on those features. We use only scores assigned to the HFT class for each observation, as the non-HFT score is merely a constant minus the HFT score. As a result, the simulated dataset consists of the same variables as the original dataset but minus the member type flags, which are replaced with the HFT scores. To simplify our explanation based on fuzzy logic, we performed a simple linear transformation of the scores to probability by scaling them to their maximum. The linear transformation ensures that linear relationships have been preserved. As a result, the transformed score represents the probability that an observation belongs to the HFT class.

The same reasons that led us to choose a decision tree for classification also led us to select a regression decision tree for our simulated data. The simulated dataset only differs in member types, which are now replaced by a continuous variable of scores. Scores are considered the dependent variable in a regression tree model.

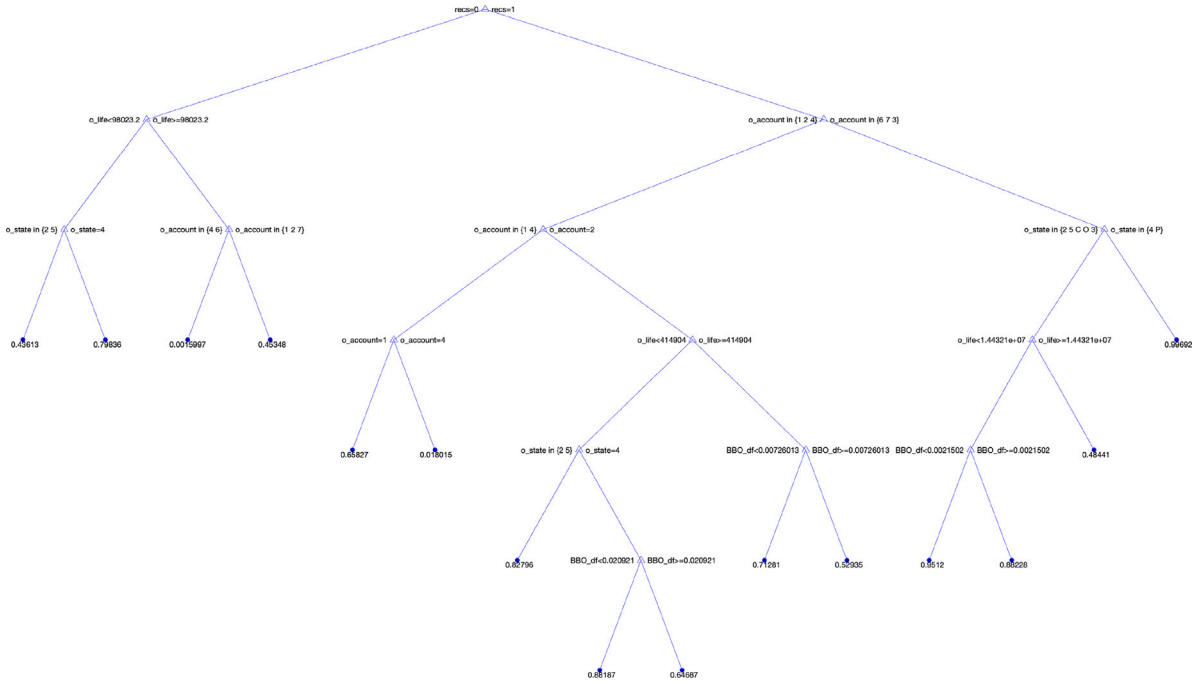

Fig. 2. Pruned tree regression of simulated data.

A regression tree model is constructed based on the simulated dataset. Overfitting is of no concern and even preferred in this classification, which is solely used to reverse engineer ensemble training. Indeed, we will build an interpretable regression tree representing the black box of ensemble trees. As a result, to achieve a perfect fit between the scores and the features, we eliminated those observations misclassified in the ensemble model, which made up about 1% of the data.

To ensure the best possible fit, we constructed a deep regression tree, and then to make an interpretable model, we pruned it back to a maximum of 15 leaves. It has long been demonstrated that building a deep tree and then pruning it to a specific number of leaves yields a more accurate model than simply building the tree with that number of leaves.

A single tree, the representative of ensemble model

Fig. 2 shows the pruned regression tree, which represents the ensemble tree boosted by the random under-sampling method. The final values in each leave show the probability of being HFT. With fifteen nodes in the decision tree, this simplified regression model offers insights into the relationship between the selected features and the probability of an order being classified as HFT. The tree structure reveals the importance of variables such as ‘‘recs’’ (trading group regulation), ‘‘order life’’, ‘‘state of the order’’, ‘‘account type’’, and ‘‘BBO_diff’’ in determining the HFT classification.

Table 8 listed these groups from the most probable states of being an HFT order to the least likely.

Only five features represent the ensemble model in the regression tree and the table below. This means that the interpretable and simplified version of the ensemble tree uses only five of the 16 features to describe a simple regression model with fifteen nodes. Two sets of data for both models are used to determine how close the regression estimate is to the estimate of the ensemble model.

Fig. 3 shows the cumulative distribution function for both in-sample and out-of-sample estimations. We used data from January 2nd 2017, for out-of-sample. As one might expect, there are fewer kinks in the simple regression tree than in the ensemble model. It may stay very close to the original model along the entire graph.

To test the predictive power of the simplified regression tree, we estimated the HFT probability for January 2nd 2017, data and then categorised them into HFT and non-HFT data. Probability less than 50% replaced with class non-HFT and others with class HFT. PPV and NPV were 97.1% and 79.7%, respectively, very close to those in the original ensemble model reported in Table 5.

Table 8 shows the sorted leaves of the regression tree from high to less probable HFT. For example, the first row says if an order placed by a market maker or parent company for equity which is being traded in a regulated market with the continuous auction and static collar (`recs`) and then is cancelled by the broker or by STP mechanism, it is an HFT order with a probability of 99%. On the contrary, if an order is placed by a Retail Market Organization (RMO) or a market maker for security, not in `recs` and remains in the order book for more than 98 s, it is a non-HFT order with a probability of 99.8%.

Unlike `recs`, account, and order state, which are limited to the categories included, the BBO difference and order life domain are continuous, and the thresholds are likely to change in another model. Yet what is significant here is to illustrate the ability of regression trees and fuzzy logic to interpret ensemble models that provide accurate classification through non-interpretable processes.

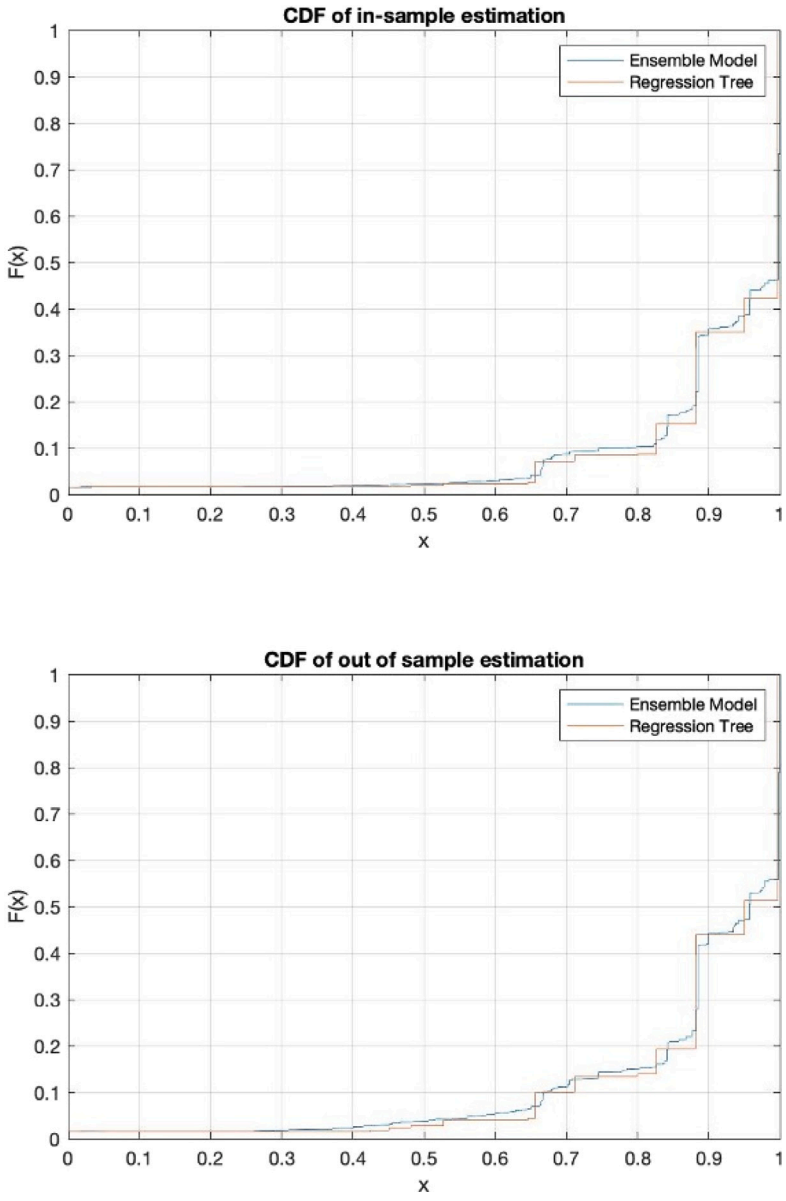

Fig. 3. Cumulative distribution function.

As can be seen from the tree graph and the sorted table, `recs` and order life are the most distinguishing characteristics, meaning that almost all orders for securities with `recs` and all orders with order life less than the specified thresholds are more likely to be HFT.

5. Discussion and conclusion

The empirical results from the evaluation of the models established that the ensemble model combined with random under-sampling among several supervised learning approaches is the most effective method for detecting HFTs in order book data. The model showed a very robust performance in classifying data from five other trading days during a week and performed very well on all those days. The method requires feature engineering, and those extracted features based on theory were among the most differentiating ones in the final model and the regression tree. For example, the shorter order life and tendency to the securities with the highest liquidity have long been attributed to HFT in the literature.

To shed light on the black box of the ensemble model, we took a reverse engineering approach to the fullest extent possible. We assigned the ensemble score to the features in an interpretable manner. As a result, we could build a regression tree that returned

Table 8
Sorted list of the states resulted from regression tree.

recs	Account	Order life	Order state	BBO Diff	Prob. of HFT	Type of order
1	6, 7, 3		4, p		99%	Order for recs placed by Market Makers (MM) or Retail MM or parent company account that is cancelled by the broker or self-trade prevention (STP) mechanism
1	6, 7, 3	< 14 389	2, 5, C, O, 3	< 0.2%	95%	Order for recs placed by MM or Retail MM or parent company account that totally filled, modified, eliminated by day validity, eliminated by corporate events or cancelled by trading system within 14 389 seconds, with a price in range of BBO ±0.2%
1	6, 7, 3	< 14 389	2, 5, C, O, 3	> 0.2%	88%	Order for recs placed by MM or Retail MM or parent company account that totally filled, modified, eliminated by day validity, eliminated by corporate events or cancelled by trading system within 14 389 seconds, with a price out of range of BBO ±0.2%
1	2	< 428	4	< 2%	88%	Order for recs placed by a proprietary account, cancelled within 428 seconds and its price is in range of BBO ±2%
1	2	< 428	2, 5		83%	Order for recs placed by a proprietary account totally filled or modified within 428 s
0		< 98	4		80%	Order for securities not in recs that has been cancelled by the broker within 98 s
1	2	> 428		< 0.7%	71%	Order for recs placed by a proprietary account, with a life more than 428 s and a price in range of BBO ±0.7%
1	1				66%	Order for recs securities placed by a client account
1	4	< 428		> 2%	65%	Order for recs placed by a proprietary account, cancelled within 428 s and its price is out of range of BBO ±2%
1	2	> 428		> 0.7%	53%	Order for recs placed by a proprietary account, with a life more than 428 s and a price out of range of BBO ±0.7%
1	6, 7, 3	> 14 389	2, 5, C, O, 3		48%	Order for recs placed by MM or Retail MM or parent company account that totally filled, modified, eliminated by day validity, eliminated by corporate events or cancelled by the trading system and with a life of more than 14 389 seconds
0	1, 2, 7	> 98			45%	Order for securities not in recs placed by clients account, proprietary account or parent company account and lasted more than 98 s
0		< 98	2, 5		43%	Order for securities not in recs, totally filled or modified less than 98 s
1	4				1.7%	Order for securities in recs placed by a RMO
0	4, 6	> 98			0.2%	Order for securities not in recs placed by Market makers (MM) or Retail Market Organizations (RMO) with an order life of more than 98 s

the probability of being HFT. The regression tree that most closely resembles the ensemble model showed very high accuracy in classification, regardless of whether in-sample data was used.

The interpretation of the regression tree and its branches goes beyond model accuracy. It offers insights into the strategies associated with HFT, revealing valuable information for policymakers and researchers interested in understanding the dynamics and behaviours of HFT participants. This interpretation contributes to the literature by supporting the theoretical definition of HFT and providing a simple and innovative method for interpreting non-interpretable classification models.

This study adds to the existing literature in several significant ways. Firstly, the resulting model demonstrates exceptional precision and reliability in detecting HFT. Moreover, the approach employed in this study can be applied to any market with similar order book data, enabling the identification of HFT in real-time, unlike existing techniques that rely on annual data. This real-time capability empowers regulators to monitor and regulate HFT activities following order completion promptly. Additionally, publicly available data, as opposed to relying solely on regulatory data, facilitates broader research on HFT and removes potential data limitations, leading to further advancements in HFT identification.

Furthermore, the methodology presented in this study has the potential to establish a standardised approach to HFT identification. Unifying the identification process can eliminate inconsistencies arising from various methods, resulting in more consistent and reliable research outcomes. This standardisation can enhance the understanding of HFT's impact on financial markets and facilitate comparative studies across different markets and regulatory environments.

While the developed model is currently tailored to stock markets with order book data similar to the AMF database, the comprehensive methodology presented can be transferable to other stock exchanges. Researchers and practitioners can utilise the step-by-step approach, encompassing data preparation, feature extraction, model selection, and interpretation, as a guideline for conducting similar studies in diverse market contexts. This transferability fosters future research on HFT detection and promotes a deeper understanding of HFT dynamics across global markets.

Additionally, the created model was interpreted in detail, down to the probabilities assigned to each branch of the decision tree, thanks to this study's novel reverse engineering approach to decode the ensemble tree's black box. Using the idea of fuzzy membership, policymakers can benefit from this probabilistic categorisation and develop flexible policies. Indeed, the interpretation of the ensemble model contributes theoretically to the literature by supporting the theoretical definition of HFT and methodologically by presenting a simple and innovative way of interpreting non-interpretable classification models. Policymakers and researchers can interpret very accurate models and thus gain an understanding of the aggregate behaviour of HFTrs, which can lead to further investigation and optimisation of policies.

While this study's results may help detect HFT using the AMF database, the methods used for feature engineering, model selection, and reverse engineering to interpret the model generally apply to any dataset. Therefore, our research not only aids in identifying HFTs in the French market but also provides methodological guidance for the broader task of identifying HFTs in general through the application of machine learning. By leveraging the strengths of the model and its methodology, policymakers and researchers can gain valuable insights into the behaviour and impact of HFT, paving the way for a better understanding and management of this increasingly prevalent trading strategy.

CRediT authorship contribution statement

Mostafa Goudarzi: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Flavio Bazzana:** Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

Data availability

The authors do not have permission to share data.

Acknowledgements

We sincerely thank one anonymous reviewer for his careful reading of our manuscript and his many insightful comments and suggestions. We acknowledge the support of the "Equipex PLADIFES ANR-21-ESRE-0036 (France 2030)".

References

Aitken, M.J., Aspris, A., Foley, S., de B. Harris, F.H., 2018. Market fairness: The poor country cousin of market efficiency. J. Bus. Ethics 147, 5–23.
AMF, 2017. Study of the behaviour of high-frequency traders on Euronext Paris. Risk Trend Mapp.
Ammar, I.B., Hellara, S., Ghadhab, I., 2020. High-frequency trading and stock liquidity: An intraday analysis. Res. Int. Bus. Finance 53, 101235.
ASIC, 2015. Review of high-frequency trading and dark liquidity. Report 452, 20.
Bazzana, F., Collini, A., 2020. How does HFT activity impact market volatility and the bid-ask spread after an exogenous shock? An empirical analysis on S&P 500 ETF. North Am. J. Econ. Finance 54, 101240.
Bellia, M., 2018. Essays on Empirical Market Microstructure and High Frequency Data. Technical Report, Università Ca'Foscari Venezia.
Benos, E., Sagade, S., 2016. Price discovery and the cross-section of high-frequency trading. J. Financial Mark. 30, 54–77.
Bernales, A., 2019. Make-take decisions under high-frequency trading competition. J. Financial Mark. 45, 1–18.
Biais, B., Foucault, T., et al., 2014. HFT and market quality. Bank. Mark. Invest. 128, 5–19.
Boehmer, E., Li, D., Saar, G., 2018. The competitive landscape of high-frequency trading firms. Rev. Financ. Stud. 31 (6), 2227–2276.
Breckenfelder, J., 2019. Competition Among High-Frequency Traders, and Market Quality. Technical Report ECB Working Paper.
Brogaard, J., Garriott, C., 2019. High-frequency trading competition. J. Financ. Quant. Anal. 54 (4), 1469–1497.
Brogaard, J., Hendershott, T., Riordan, R., 2014. High-frequency trading and price discovery. Rev. Financ. Stud. 27 (8), 2267–2306.
Brogaard, J., Hendershott, T., Riordan, R., 2017. High frequency trading and the 2008 short-sale ban. J. Financ. Econ. 124 (1), 22–42.
Comerton-Forde, C., Malinova, K., Park, A., 2018. Regulating dark trading: Order flow segmentation and market quality. J. Financ. Econ. 130 (2), 347–366.
Ekinci, C., Ersan, O., 2018. A new approach for detecting high-frequency trading from order and trade data. Finance Res. Lett. 24, 313–320.
Ekinci, C., Ersan, O., 2022. High-frequency trading and market quality: The case of a "slightly exposed" market. Int. Rev. Financ. Anal. 79, 102004.
Elizarov, M., Ivanyuk, V., Soloviev, V., Tsvirkun, A., 2017. Identification of high-frequency traders using fuzzy logic methods. In: Tenth International Conference Management of Large-Scale System Development. MLSD, IEEE, pp. 1–4.
Ersan, O., Ekinci, C., 2016. Algorithmic and high-frequency trading in Borsa Istanbul. Borsa Istanbul Rev. 16 (4), 233–248.
Hagströmer, B., Nordén, L., 2013. The diversity of high-frequency traders. J. Financial Mark. 16 (4), 741–770.
Hagströmer, B., Nordén, L., Zhang, D., 2014. How aggressive are high-frequency traders? Financ. Rev. 49 (2), 395–419.
Han, H., Forrest, J.Y.L., Wang, J., Yuan, S., Fei, H., Li, D., 2022. Explainable machine learning for high-frequency trading dynamics discovery. Available at SSRN 4256777.
Hasbrouck, J., Saar, G., 2013. Low-latency trading. J. Financial Mark. 16 (4), 646–679.
Hossain, S., 2022. High-Frequency Trading (HFT) and Market Quality Research: An Evaluation of the Alternative HFT Proxies. J. Risk Financial Manag. 15 (2), 54.
Jarnecic, E., Snape, M., 2014. The provision of liquidity by high-frequency participants. Financ. Rev. 49 (2), 371–394.
Kang, J., Kang, J., Kwon, K.Y., 2022. Market versus limit orders of speculative high-frequency traders and price discovery. Res. Int. Bus. Finance 63, 101794.
Kelejian, H.H., Mukerji, P., 2016. Does high frequency algorithmic trading matter for non-at investors? Res. Int. Bus. Finance 37, 78–92.
Kirilenko, A., Kyle, A.S., Samadi, M., Tuzun, T., 2017. The flash crash: High-frequency trading in an electronic market. J. Finance 72 (3), 967–998.
Korajczyk, R.A., Murphy, D., 2019. Do High-Frequency Traders Improve your Implementation Shortfall? J. Invest. Manag. 18, 18–33.
Leone, V., Kwabi, F., 2019. High frequency trading, price discovery and market efficiency in the FTSE100. Econom. Lett. 181, 174–177.
Li, K., 2021. Does high-frequency trading impede order execution in the stock market? Procedia Comput. Sci. 187, 501–506.
Malinova, K., Park, A., 2020. 'Sniping' in fragmented markets. Available at SSRN 3534367.
Mankad, S., Michailidis, G., Kirilenko, A., 2013. Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method. Algorithmic Finance 2 (2), 151–165.
Menkveld, A.J., 2016. The economics of high-frequency trading: Taking stock. Annu. Rev. Finan. Econ. 8, 1–24.
O'Hara, M., Saar, G., Zhong, Z., 2019. Relative tick size and the trading environment. Rev. Asset Pricing Stud. 9 (1), 47–90.
Scholtus, M., Van Dijk, D., Frijns, B., 2014. Speed, algorithmic trading, and market quality around macroeconomic news announcements. J. Bank. Financ. 38, 89–105.
SEC, 2010. Concept release on equity market structure. Fed. Regist. 75 (13), 3594–3614.
Van Ness, B.F., Van Ness, R.A., Watson, E.D., 2015. Canceling liquidity. J. Financial Res. 38 (1), 3–33.
Van Vliet, B., 2017. Capability satisficing in high frequency trading. Res. Int. Bus. Finance 42, 509–521.
Yang, H., Ge, H., Luo, Y., 2020. The optimal bid-ask price strategies of high-frequency trading and the effect on market liquidity. Res. Int. Bus. Finance 53, 101194.