IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 18, 2021, accepted November 6, 2021, date of publication November 25, 2021, date of current version December 9, 2021.

*Digital Object Identifier 10.1109/ACCESS.2021.3130843*

# Machine Learning in Financial Market Surveillance: A Survey

**SHWETA TIWARI, HERI RAMAMPIARO, AND HELGE LANGSETH**
Department of Computer Science, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway

Corresponding author: Shweta Tiwari (shweta.tiwari@ntnu.no)

This work was supported by the Department of Computer Science, the Norwegian University of Science and Technology, and the Norwegian Open AI Lab. In addition, the ELSPOT data used in this paper was provided by NordPool AS, in collaboration with Optimeering AS.

**ABSTRACT** The use of machine learning for anomaly detection is a well-studied topic within various application domains. However, the detection problem for market surveillance remains challenging due to the lack of labeled data and the nature of anomalous behaviors, which are often contextual and spread over a sequence of anomalous instances. This paper provides a comprehensive review of state-of-the-art machine learning methods used, particularly in financial market surveillance. We discuss the research challenges and progress in this field, mainly applied in other related application domains. In particular, we present a case of machine learning-based surveillance system design for a physical power trading market and discuss how the nature of input data affects the effectiveness of the methods on detecting anomalous market behaviors. Overall, our findings indicate that the regression tree-based ensemble algorithms robustly and effectively predict day-ahead future prices, showing their capability to detect abnormal price changes.

**INDEX TERMS** Anomaly detection, financial market surveillance, machine learning, time series.

## I. INTRODUCTION

Financial markets allow traders to profit from buying and selling financial instruments without physically owning the underlying assets. Although a trader can own an asset for a longer time, such as in options markets, prices of assets in a market generally rise or fall due to short selling, attracting more people to speculate merely on price movements. A perfect market is both liquid and efficient so that there are no sudden volatile changes in the prices and volumes of traded instruments, i.e. a relatively stable market [1]. However, in practice, both price and volume of an instrument in a financial market can be influenced by undisclosed information associated with the instrument and several trade-based manipulations [2].

Market abuses are subjective and are generally decided by regulations and the guidelines governing the market. Nevertheless, although there is no standard definition of market abuse, there are two main categories of abuses that have been widely studied in the literature [2], [3]. The first category consists of practices that result in illegal benefits from the market by strategically manipulating market prices [2], [4], [5]. There are several types of activities that commonly fall under the category of price manipulation, including wash trade, spoofing, and late-day trading. The price manipulation strategies that involve influencing prices in the short term have a common factor: the essential information upon which an instrument's price is based cannot be changed in the long term. Price manipulation remains the biggest concern for the free and fair functioning of financial markets. The second category of market abuse, known as insider trading [6], occurs due to revealing classified information by an insider. The insider information can affect market prices, and market actors can exploit this to make windfall gains.

All trading activities in financial markets today happen on electronic platforms. Due to readily available information and ease of conducting trading activities using automated means, greater demands for market transparency have led every marketplace to lay down specific regulatory requirements for traders and market actors to comply with. Today, regulatory surveillance of financial markets is mainly based on trading data used to study market dynamics and analyze suspicious trading activities, such as price manipulation and insider trading. However, the huge amount of data poses challenges for its storage, processing, and analysis. In order to cope with

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

the large data volumes, rule-based approaches have been used to identify potentially suspicious market events that meet a pre-specified set of conditions (rules) [7], [8]. Events that trigger automatic ''warnings'' or ''alerts'' are typically passed on to human experts to analyze and process manually within the surveillance workflow, e.g., [9]. However, a rule-based system typically generates a large volume of alerts, especially as the complexity and volume of transactions increase. In addition, it is time-consuming, difficult, and costly to determine these alerts as true positives or not. For example, in a rapidly growing physical power market, several trading activities, such as spot, multiple reserves, and intra-day trading incorporate additional complexity in monitoring. However, a rule-based surveillance system monitors only those events for which it has been designed, and therefore new market developments and manipulation strategies or approaches may quickly make it obsolete.

### A. KEY RESEARCH CHALLENGES

At an abstract level, the task of financial market surveillance is to define a region representing normal behavior and classify any activity or observation in the data which does not belong to this normal region as a manipulation. However, the task of algorithmically learning to categorize abnormal and normal situations from market data automatically is non-trivial and involves several real-world challenges.

First, anomalies are by definition rare. Hence, the amount of labeled data that can be used for training a machine to recognize such events is very small and labeled data are costly to produce. Second, separating anomalous and normal instances requires defining a boundary that encompasses every possible normal behavior. However, this boundary is often not precise, and therefore data instances close to the boundary often get misclassified. In anomaly detection algorithms, models are first trained to compute scores of each data point, and then data instances that receive the highest scores are reported as anomalies [10]. Human analysts then identify true anomalies by analyzing top-ranked anomalies. However, many true anomalies reported by anomaly detection algorithms could be false-positives resulting from data instances that did not fit a normal model. Third, data often accumulate noise due to variability involved in its generation, collection and processing, which further complicates detecting real anomalies, often generating more false-positives. Fourth, an abnormal market behavior event may not simply be a single market action, but a series of market actions carried out by an actor. The sequence of actions, time and order between them matters. Hence the sequential information incorporated in the order and period of market actions must be captured and considered by the system to categorize market behavior as an anomaly [11].

The first three research challenges mentioned above are common within the anomaly detection problem in any application area. This paper provides an in-depth description of anomaly detection methods proposed over time for addressing these challenges in the market surveillance design and other related application areas. The fourth research challenge is specifically related to designing a machine learning-based system that learns patterns from time-series data, makes predictions and detects anomalies. We discuss the core elements of this problem, provide a review of research done in the area and discuss technical challenges in the context of machine learning-based surveillance system design. In particular, we did a comparative study on different machine learning (ML) methods for learning patterns in time-series datasets, making predictions, and identifying anomalies. In addition to discussing the results from this study, we present findings concerning learning patterns, effectively predicting day-ahead future prices and generating abnormal price changes in time-series data from the electricity trading market.

### B. OTHER RELATED SURVEYS

Financial markets are witnessing changes at an unprecedented speed. Following this, the regulations that govern these markets also change quickly, thus requiring the regulatory surveillance mechanisms to adapt continuously. Machine learning-based surveillance of financial markets has not started to gain prominence in research until very recently after the emergence of scalable and adaptive methods in AI and machine learning [12]. To the best of our knowledge, there does not exist other surveys that directly discuss machine learning-based financial market surveillance, particularly concerning commodities markets like the power trading market. For this reason, related surveys discussed in this section cover other fields closely related to the context of this paper.

In their work, Ahmed *et al.* [13] presented a review of unsupervised machine learning methods used for fraud detection in the financial domain. Hodge and Austin [14] provided a systematic survey of machine learning techniques and statistical methods used in anomaly detection and discussed their advantages and disadvantages with respect to the application domains. Chandola *et al.* [10] gave a broad and structured overview of anomaly detection techniques in different application domains and research areas. Akoglu *et al.* [15] provided an extensive review on graph dataset and graph-based anomaly detection. In the most recent reviews, Chalapathy and Chawla [16], Pang *et al.* [17] focused on deep anomaly detection and presented a survey covering the state-of-the-art deep neural network-based techniques.

Note that since the main focus of this paper is to give a thorough literature review of machine learning methods used in financial market surveillance and other related application domains, a detailed review of graph-based methods and statistical methods is considered beyond the scope of this paper. Nevertheless, we provide a brief overview of statistical methods for anomaly detection in Section III. The reader may refer to [15] and [10] for detailed reviews of anomaly detection using graph-based methods and statistical methods, respectively.

### C. MAIN CONTRIBUTIONS

This paper has three main contributions. First, we provide a comprehensive literature review of machine learning methods specifically designed for financial markets monitoring and surveillance and other related application domains. Second, we perform an extensive comparative study on state-of-the-art machine learning methods used for detecting anomalies in time-series data of electricity prices, which, although being used in related applications, were previously applied in other domains. As a case of the emerging financial market for trade surveillance design using machine learning methods, we present NordPool's physical power market.[1] Third, to provide more in-depth knowledge of relevant methods, we systematically evaluate the methods using two different datasets, including electricity price data obtained from NordPool[2] and the Numenta Anomaly Benchmark (NAB) [18] dataset. The results from the experimental evaluation using commonly applied metrics for anomaly detection show the advantages and weaknesses of existing methods when applied to detect anomalies in time-series data. To the best of our knowledge, no other studies directly survey machine learning methods for trade surveillance within electricity markets. In summary, this paper gives the reader an extensive overview of existing methods and provides a comparative study with experimental evaluation.

## II. BACKGROUND

In this section, we start our discussion with a representative example of the financial market, Europe's leading physical power market that NordPool conducts, and its surveillance to detect abnormal market activities. Further, we illustrate the main building blocks of a market surveillance system. We then give a general background of anomalies and their taxonomy in the context of machine learning. At the end of the section, we present an ML-based pipeline of the analytical engine in the trading surveillance system.

### A. ELECTRICITY TRADING MARKET SURVEILLANCE

Financial markets refer broadly to any marketplace where securities are traded by buying and selling financial instruments such as equities, bonds, currencies, and derivatives. Market surveillance is typically based on monitoring trading activities to detect or predict potential abusive behaviors of market actors. The monitoring is primarily based on analyses of historical trading data. For example, short-term prices of a market instrument can be predicted using forecasting approaches based on time-series methods. The forecasting model can be developed using historical data and parameter tuning, independent from the underlying market structure.

To discuss the general terminology of a financial market and its surveillance mechanism, we consider the example of NordPool exchange. NordPool is one of the largest electricity trading markets[3] in the world, measured in the volume traded (a total of 524 TWh in 2018). It operates in the Nordic region (Sweden, Norway, Denmark, Finland), the Baltic region (Estonia and Lithuania), Germany, France, Netherlands, Belgium, Austria, and the UK. More than 90% of the total electricity consumption in the Nordic region is traded through NordPool. Electricity market players, such as electricity producing firms, large consumers, distributors, banks, brokers, and others (today 380 different participants from 20 different countries trade on NordPool) trade electricity on NordPool's spot market [19], which is divided into two sub-markets: Elspot and Elbas. In the Elspot market, buy and sell orders are placed by traders on an hourly basis for physical delivery the next day. The Elbas market, on the other hand, conducts intra-day trading that involves cross-border delivery of physical power, where traders can place orders until one hour prior to delivery, and the trade adjustments are made in the day-ahead market. The overall system price is calculated based on the equilibrium between the aggregated demand and supply generated by all buy and sell orders.

Since the beginning of their commercialization, the power markets have been monitored to ensure market efficiency. In the early stages, the greatest concern was whether a participant was taking advantage of his market power. Later, market monitors have learned that the power market is susceptible to more diverse market manipulation strategies [20]. The abusive behaviors of market participants are categorized into two main classes: exercising market power and market manipulation strategies. The former usually refers to the act of physically or economically withholding capacity. For market manipulation, the European Union (EU) Regulation on Wholesale Energy Market Integrity and Transparency (REMIT) has defined four categories of market manipulation or types of attempts that amount to market manipulation:

- False or misleading orders/transactions, e.g. wash trades, marking the close, cross-market manipulation and spoofing.
- Price positioning such as pre-arranged trading.
- Fictitious device or deception.
- Dissemination of false or misleading information.

Most descriptions of manipulation refer to the first two categories. Examples are layering and spoofing, marking the close, pre-arranged trading, wash trades and capacity hoarding.[4] In layering and spoofing, fake bids are issued to send misleading signals to other market participants with the goal of them changing their prices. Shortly after the desired price change, or transaction, the fake bids are cancelled. The fake bids are issued on one side of the order book to obtain a more favorable price on the other side of the order book. For example, in selling energy, one would issue orders for buying energy which is generally lower than all other bids. Marking the close refers to an activity and transactions that

[1] https://www.nordpoolgroup.com/

[2] https://www.nordpoolgroup.com/Market-data1/Dayahead/Area-Prices/ALL1/Hourly/

[3] https://www.nordpoolgroup.com/trading/Day-ahead-trading/

[4] https://www.emissions-euets.com/market-manipulation-remit/

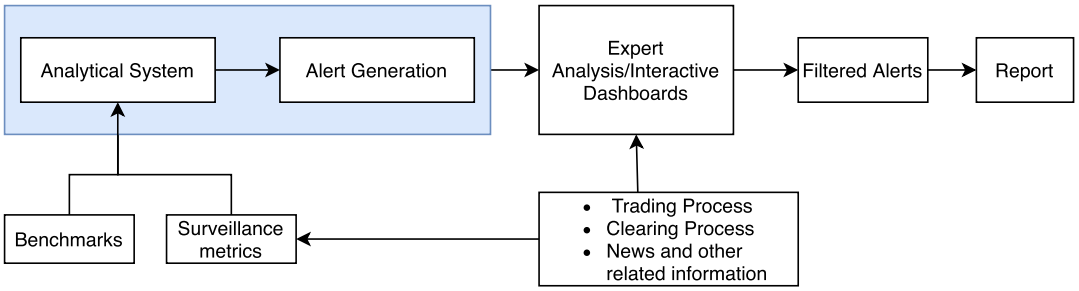



**FIGURE 1.** Overview of the surveillance process.

happen just before the market is closing. Trading at market closure gives other market participants no time to react to the changes, and the closing price could be locked at an artificial level. Pre-arranged trading is defined as market manipulation, as it may give misleading information to the market. The price and quantity have been agreed upon terms which are unknown to the rest of the market. Wash trades refer to transactions where buyer and seller are the same people or buyer and seller work in collusion. Such transactions give false information about the real market risk and interests of the market participants. The last strategy mentioned above, capacity hoarding, is a form of market manipulation where a market participant acquires a predetermined part of the available capacity with an inefficient use or without using it at all.

Figure 1 illustrates a general pipeline of a trading surveillance system [21]. As shown in this figure, the main components of the system are (1) market data, (2) surveillance metrics, (3) analytic engine and alert generation, and (4) human expert who decides on final alerts that are investigated and reported. Market data comes from different market activities, such as the trading process, electronic communications, the social media information of market actors, news feeds, and so on. It is broadly represented along two dimensions, structured/unstructured and historical/real-time. The massive amounts of market data are selectively analyzed to generate tip-offs signaling suspicious market behaviors based on previously observed similar misconduct behaviors. In order to decide what to measure in the market data to generate tip-offs, a typical strategy is to define specific metrics. For example, observed market prices could serve as a metric to detect price manipulation in the electricity trading market. Then, this metric can be used to compare the observed prices against existing benchmarks that define normal and abnormal prices.

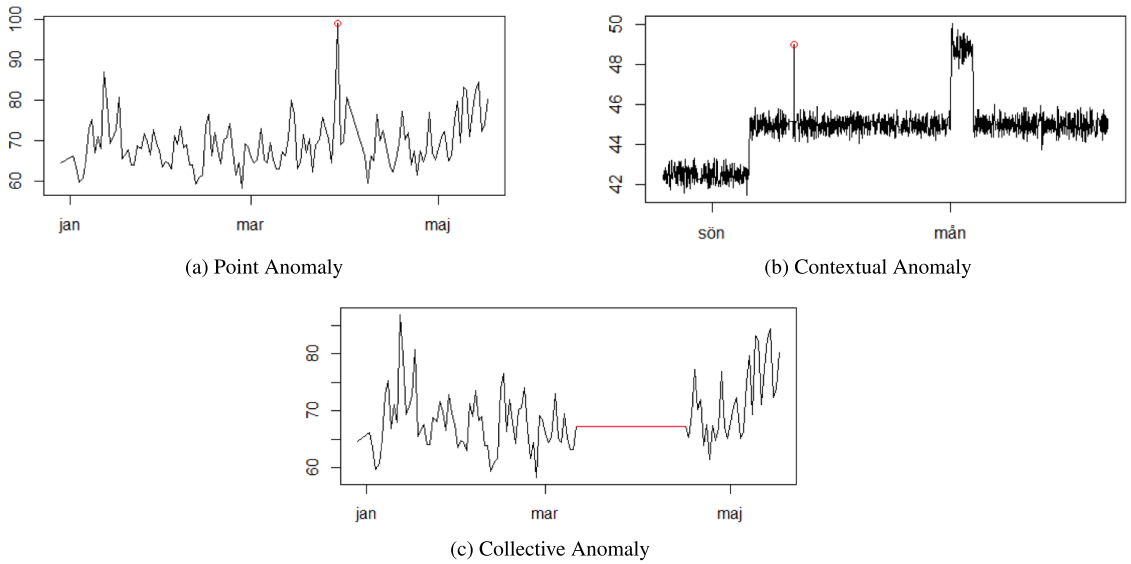
Once we have chosen data and surveillance metrics for the analysis, the next step is to use machine learning methods (see Figure 3 for a typical pipeline of ML-based analytic system and alert generation) to detect anomalies. There are generally a large number of anomalies detected, but not all of them are true positives. Therefore, these sequences of the detected anomalies require further analysis to reduce the number of false positives before a sizable number of potential anomalies are selected to generate alerts that are passed on to experts for further analysis. Since these alerts signal suspicious transactions, experts need to understand the intention behind those trades. Experts then can use other forms of analyses and domain knowledge to decide which alerts ultimately need to be reported and investigated.

### B. ANOMALY DETECTION

Anomaly detection (AD) is a widely studied topic in the field of data mining [22], statistics, and machine learning [13], [14]. It has emerged as an important tool in finding abnormalities in various domains, such as credit card fraud detection [23], financial transaction fraud detection [24], cyber intrusion detection [25], [26], and so on. Anomaly or outlier detection refers to a problem of detecting data points and/or patterns representing behaviors and/or events that deviate significantly from those considered normal data. These points are anomalies, outliers, discordant observations, exceptions, aberration, surprises, peculiarities or contaminates in different application domains. Often in the literature, anomaly detection, outlier detection, and novelty detection are used as interchangeable synonyms [27]

In a variety of the domains mentioned above, data is collected as time series. In time-series data, time is an independent variable, and physical quantities measured against it are dependent variables. In recent years, researchers have been increasingly interested in analyzing unusual but interesting phenomena in time-series data. Financial marker fraud detection is one of the examples where time-series data from market activities is used to detect and analyze outliers; often, these observations are referred to as anomalies [28].

Anomalies mainly consist of three categories: collective, contextual, and point [10]. Point anomalies are data instances significantly deviating from the other data instances. Figure 2(a) shows an example of point anomalies in time-series data from a temperature sensor. Most of the temperature values range between 60 to 90, a single instance value near 100 looks anomalous as it deviates significantly from the other data points. Contextual or conditional anomalies [29] refer to a single data point or a set of anomalous data points concerning its local neighborhood but not otherwise. Figure 2(b) shows an example of a contextual anomaly in time-series data. Although there are two peaks of the almost




(a) Point Anomaly

(b) Contextual Anomaly

(c) Collective Anomaly

**FIGURE 2.** Types of anomaly in time series.

same amplitude, one between Sunday and Monday and the other early on Monday, the data instance regarding the former represents an abnormal behavior while considering its local neighborhood. Collective anomalies represent a continuous set of data instances that are collectively anomalous even though the single instances may not necessarily be point anomalies. Figure 2 (c) shows time-series data where data instances (highlighted region) have the same value over an unusually long time span, and therefore they represent a case of collective anomalies.

The basic task in anomaly detection is to separate normal behavior from abnormal behavior. However, anomalous data instances generally constitute only a tiny fraction of all data instances. Therefore, contrary to a typical classification problem, where both the classes are expected to be balanced, an anomaly detection task represents a class imbalance problem [30]. Anomaly detection algorithms work on an alternative approach where models are first trained to compute scores of each data point, and then data instances that receive the highest scores are reported as anomalies [31]. However, many anomalies reported by anomaly detection algorithms could be false positives resulting from data instances that did not fit a normal model. In addition, data often accumulate noise due to variability involved in its generation, collection and processing, which further complicates the problem of detecting real anomalies, often generating more false positives. Most anomaly detection techniques are domain-specific. For example, techniques developed specifically for credit card fraud detection may not detect anomalies in the stock market [32]. In some domains, normal behavior continuously evolves, such that a current specification of normal behavior may not be valid for future anomaly detection.

In financial markets, an abnormal market behavior event is not simply one single market action, but rather a series of market actions carried out by an actor. The sequence of actions and the time span and order between them matters. Hence, the sequential information incorporated in the order and time span of market actions must be captured and considered by the system in order to categorize market behavior as an anomaly [11]. From the categorization of anomalies by Chandola *et al.* [10], an abnormal market behavior would be more often a contextual and/or collective anomaly than a point anomaly [33]. Market surveillance requires proactively detecting such manipulative market behaviors and taking timely action to prevent them from disrupting the smooth functioning of the market. Machine learning provides methods for learning patterns/behaviors in historical data and predicting future events based on experience. Some of the potential benefits of using machine learning-based market surveillance are: (1) it works better than traditional rule-based systems in an evolving market, (2) it improves the quality of alerts, and (3) it can adapt to new data and handle large datasets.

Hodge *et al.* [14] categorize anomaly detection methods in three main types based on underlying machine learning approaches. First, supervised anomaly detection (SAD), where training data consists of normal and abnormal instances, is used to build a model. The trained model is then used to classify the unlabeled instances of the test dataset. Second, unsupervised anomaly detection (UAD), where the model is built using unlabeled data, assuming that anomalies are separate from the normal data. The UAD assumes that a normal pattern occurs far more frequently than the anomalous patterns. Third, semi-supervised anomaly detection (SSAD) techniques are used for building models where data is only partially labeled.

*Anomaly Detection Pipeline:* Figure 3 illustrates a generic anomaly detection pipeline, with an underlying machine learning-based surveillance system proposed in the literature [34]. The pipeline mainly consists of three modules:

1) data preprocessing, 2) learning approach, 3) testing and anomaly detection module. We briefly discuss each module in the following.

1) Data preprocessing: Input to the pipeline is raw data, which is often messy and, therefore, needs cleaning and preprocessing before it can be used to train and test machine learning models. The nature of data plays a crucial role in deciding which anomaly detection methods should be used. For example, time-series data is collected at different time intervals, where each data instance has the same number of attributes or features. Attribute values could be numerical, categorical, or complex such as images, videos etc. Real-world data is often mixed with different types of attribute values and contains missing values. For example, data imputation techniques [35] can be applied to the raw data; if the data has missing values and/or if the data contains a mix of categorical and numerical attributes, the dummy variables can be generated to convert categorical attributes into numerical attributes. Once the data is preprocessed, it is easy to extract relevant features by applying feature extraction and selection techniques [36]. A good set of features in machine learning leads to a well-trained model for a given problem. Therefore, although not trivial, extracting relevant features is important. The data is further split into train, test and validation sets on which machine learning methods are trained and tested.
2) Learning approach: The centerpiece of an anomaly detection pipeline is the underlying machine learning model. The choice of machine learning model depends mainly on two factors (see Section III for a detailed categorization): 1) availability of labeled data, and 2) nature of input data. If there are well-annotated class labels in a given data, supervised machine learning methods work well for the task of anomaly detection, where it can be viewed as a classification problem. However, unsupervised machine learning methods are used for an anomaly detection task if the given data is fully unlabeled. There are two types of anomaly detection methods: 1) models built to calculate anomaly scores (e.g., Isolation Forest [37]), and 2) anomaly scores are directly calculated on input data without building any models (e.g., k-nearest neighbor [38]). There is a third category between the labeled and unlabeled categories, where data has only a few labels, and semi-supervised learning methods are used. Semi-supervised anomaly detection methods are of two types: 1) models trained with only normal labels (e.g., One-class Random Forest [39]), and 2) models built with unlabeled data and a few instances of labeled data (e.g., [40])
3) Testing and anomaly detection: This is the third and last module in the pipeline. Once we have a well-trained model, it is tested on the test dataset to evaluate its performance. The next step after the evaluation is to detect anomalies using the trained model. We can choose an optimum set of rules or threshold to generate anomalies based on the underlying machine learning method in the model. If the dataset has labels, a trained model can be evaluated using evaluation measures. However, if there are no labels in the data, evaluating models becomes challenging, often requiring experts to gradually add a few labels in the data through a feedback loop. Expert-in-the-loop anomaly detection or active anomaly detection [40] is an active area of research in the anomaly detection field.

### C. EVALUATION MEASURES

Since anomalies constitute a very small fraction of data instances in a given dataset that has a majority of data instances in the normal class, the task of anomaly detection faces the problem of classifying an imbalanced class. It is generally not straightforward to evaluate the performance of anomaly detection methods applied in time-series. In the case of unsupervised methods, this is even more challenging due to the lack of actual labels. The anomalies are ranked in high-to-low relevance, where a high anomaly score means a high degree of abnormality. This is in contrast to a simple performance measure based on accuracy or precision/recall. For example, if a large dataset contains 10 anomalies ranked in the top 15 outliers, this may still be a good result. The selection of an optimum threshold is an important and non-trivial task in evaluating anomaly detection methods. If a threshold is too large, then the system may miss some real anomalies. However, there is a great chance to end up with many false positives if it is too low.

*Precision* and *Recall* [28] are the standard metrics used for performance evaluation of time-series anomaly detection methods. *Precision*, represents the number of real anomalies out of all detected anomalies, whereas, *Recall*, which is also known as *Sensitivity*, is the fraction of all real anomalies that are successfully detected. There are a few other related terms, such as False Positive Rate (*FPR*), Area Under Curve (*AUC*), *Jaccard*, and $F$-scores that are used as performance metrics for anomaly detection methods. For example, the *AUC* measures to what extent an ''anomalous'' data point get a higher score than a ''normal'' data point and is particularly used as a metric for evaluating the performance of unsupervised anomaly detection methods [41]. The reader is referred to [28] for a detailed description of these metrics. We use these metrics to evaluate anomaly detection methods in our experimental study in Section IV. There are also a few recent studies that introduce new metrics to evaluate time-series anomaly detection methods for real-time applications [42] and range-based anomalies [43] that occur over a time window.

## III. MACHINE LEARNING METHODS FOR MARKET SURVEILLANCE

Price manipulation is one of the main abuses in the financial markets [2], where the manipulated target is the bid price

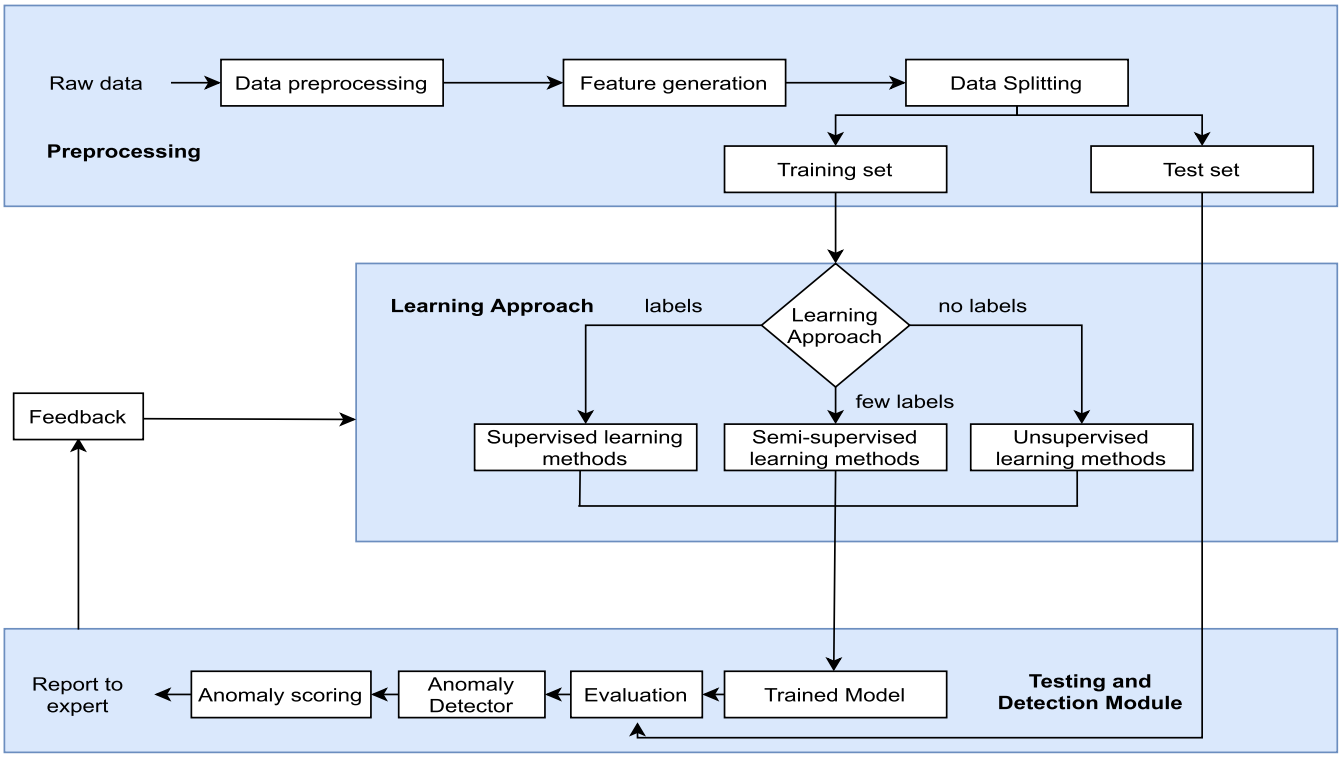



**FIGURE 3.** A generic anomaly detection pipeline underlying a machine learning-based market surveillance system.

of certain financial instruments. Price manipulation activities affect price fluctuation in capital markets where liquidity and returns unexpectedly rise and fall during the manipulation periods. Since machine learning within financial market surveillance is relatively new, scientific literature in this area is minimal. Furthermore, few studies address ML-based surveillance system design issues in emerging financial markets, such as the power trading market. Currently, there are only a few commercially available ML-based surveillance systems, such as NASDAQ SMARTS,[5] NICE Actimize,[6] Trading Technologies[7] and Scila.[8]

The Scientific literature covering machine learning methods in market surveillance can broadly be categorized based on different aspects of data. First, since anomalies are inherently rare, the amount of labeled data that can be used for training a machine learning method to detect such events are scarce and costly to produce. Therefore, selecting a machine learning technique is generally based on the availability of labeled data. Second, unusual market activities must be seen in the context of the underlying structure of the market, which like the unusual behavior, cannot be defined simply by providing data describing a snapshot of the context. Instead, the market's sequential nature must be captured to describe the context and explain the rationality behind the examined market actions. Further, within a sequence of actions, the time spans and the order between the actions matter, and the sequential information incorporated in the order and time span of market actions must be captured and considered by the machine learning algorithm to categorize market behavior as unusual [11].

In machine learning and statistics, anomaly detection has been an active area of research. Going back to Holt-Winters [44], classic and seasonal Auto-Regressive Integrated Moving Average (ARIMA) model [45], clustering techniques for detecting anomalies in time series and other types of data have been studied over the years [10], [46]. In financial market analysis and surveillance, mostly used models and tools are based on sets of rules used to define market abuse scenarios. Such rule-based systems raise an alert whenever data meets the pre-defined rules and threshold, prompting surveillance staff to investigate. SMARTS [12] is one such tool in commodity markets that uses rules to trigger alerts. Recently, machine learning techniques have gained popularity in the field of anomaly detection.

Time dependency is another crucial issue in many financial applications, such as risk management and asset allocation, because if return distribution is time-dependent, then statistical tests using unconditional statistics and inferences derived thereof could be misleading. If the time dependency can be fully exploited, it will help to produce better forecasts of level, volatility, and higher return moments [47]. The problem of detecting anomalies in time series has gained popularity in recent years. Existing research on time series anomaly

[5]http://www.nasdaq.com
[6]https://www.niceactimize.com/financial-markets-compliance/
[7]https://www.tradingtechnologies.com/trading/tt-platform/
[8]https://scila.se/

detection is fragmented across different application domains; hence it is essential to provide an overview of anomaly detection techniques for time series that cover multiple research areas and application domains. For this reason, in this section, we will discuss various proposed methods to capture different types of anomalies in the financial domain and in time-series data in general. We categorize the papers broadly based on the availability of labeled data and the nature of input data.

### A. THE AVAILABILITY OF LABELED DATA

Anomaly detection methods can be broadly classified into three categories based on the availability of labeled data: supervised models, semi-supervised methods, and unsupervised methods.

#### 1) SUPERVISED ANOMALY DETECTION

The first category includes supervised machine learning approaches that are used when labeled data with normal and anomalous instances are available for training classifiers that predict anomalies. To understand the main concept behind the supervised anomaly detection, let us consider a set of $N$ training examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ such that $x_n$ is a feature vector of the $n_{th}$ example and $y_n$ is its label then the supervised learning algorithm seeks a mapping function from $X$ to $Y$, i.e. $Y = f(X)$, where $X$ is the input space and $Y$ is the output space. Since supervised anomaly detection is similar to the standard classification setup, a variety of classification methods with good empirical performance can be used.

Golmohammadi *et al.* [32] used supervised learning methods to detect unusual transactions traders use for manipulating the stock market. Market manipulations such as marking the close, wash trade and cornering the market, as mentioned in Section II, are associated mainly with commodity price. Thus, it is essential to monitor price percentage changes. These types of manipulations can be identified by using classical methods, such as decision trees (DT), Random Forest (RF), Naive Bayes, Neural Networks (NN), k-Nearest Neighbor (k-NN), and Support Vector Machines (SVM).

The existing top-down approach for detecting financial market manipulations is based on a set of predefined thresholds and known patterns. Such an approach suffers from a large number of false positives and is usually non-adaptive. As listed in Table 1, several supervised machine learning techniques have been used to identify suspicious transactions concerning stock market manipulation. Logistic regression (LR) with an Artificial Neural Network (ANN) and SVM have been studied and compared to detect trade-based manipulation within the emerging Istanbul Stock Market [6]. Supervised methods, such as ANN and SVM, give better performance than statistical techniques regarding total classification accuracy and sensitivity (recall).

#### 2) SEMI-SUPERVISED ANOMALY DETECTION

Semi-supervised learning is a sub-class of supervised learning techniques. Training data is available only for one class in the semi-supervised anomaly detection setting, i.e., the normal class. Hence, it is more widely used than the supervised learning setting. The new data is compared against the known normal class. Then, the data that does not satisfy the normal class condition is considered an anomaly. This approach may perform well for a specific time frame, but it is prone to perform poorly if the manipulation characteristics evolve. This type of learning method is also termed novelty detection.

The novelty detection based k-NN method is one of the simplest non-parametric approaches. It computes the distance (e.g., Euclidean distance) between training samples, and then the test data uses the lowest distance score plus a threshold to make the novelty detection. One-class Support Vector Machine (OCSVM) is another approach for detecting novelty, as it provides a direct description of normality boundary [52]. A model based on k-NN and OCSVM along with a data transformation method to make non-stationary data into stationary data was proposed in [3]. The method is applied to detect spoofing and quote stuffing manipulations in stock market data.

#### 3) UNSUPERVISED ANOMALY DETECTION

Since the availability of labeled data for anomaly detection is rare and costly to produce, and the data characteristics often evolve, making it challenging to apply supervised learning methods on such data. When there is a lack of labels in the data, unsupervised anomaly detection methods can be used to detect anomalies; thus, this type of methods are most widely applicable. Many semi-supervised techniques can be adapted to operate in an unsupervised mode using a sample of the unlabeled dataset as training data. Such adaptation assumes that the train data contains very few anomalies and the model learnt during training is robust to these few anomalies. Table 1 gives an overview of the state-of-the-art unsupervised machine learning techniques applied to detect financial market manipulations.

Clustering is an unsupervised learning technique that builds different clusters of a given dataset during the training phase using some criteria. The key assumptions in clustering-based anomaly detection are as the following. First, any new data instance which does not fit in any cluster is considered anomalous. For example, density-based clustering [53] does not include noise inside the clusters. Second, when a cluster contains both normal and anomalous data instances, the normal data instances lie close to the centroid. In contrast, anomalies are those that are far from the centroids [54]. Third, if there are clusters of various sizes, larger clusters are considered normal, whereas sparse clusters can be considered anomalous [55]. Ahmed *et al.* [13] provide a detailed review of clustering-based anomaly detection techniques used in the financial domain. In modern financial markets, large amounts of data are available to market participants to gain valuable insights and make better decisions. Various clustering-based methods, such as Local Outlier Factor (LOF) and Connectivity Outlier Factor (COF), along with statistical methods, have been used to detect anomalies in

**TABLE 1.** State-of-the-art supervised, semi-supervised, and unsupervised machine learning techniques used for financial market surveillance.

| Paper | Methods | Manipulation type | Evaluation measures | Data |
|---|---|---|---|---|
| Golmohammadi *et al.* [32] | k-NN, SVM, NN, RF, DT | Wash trade | Precision, Recall, F2 measure | Diaz *et al.* [7] |
| Frery *et al.* [48] | Learning to rank | – | Average, Precision | A highly unbalanced data |
| Öğüt *et al.* [6] | SVM, k-NN, ANN, LR | Trade-based manipulation | Recall, Precision | Istanbul Stock Market data |
| Ahmed *et al.* [49], Ahmed *et al.* [13] | Clustering algorithms | Trade-based manipulation | Recall, Precision, and F2 measure | ASX data |
| Das *et al.* [40], Das *et al.* [50] | AAD, iForest-AAD, Tree-based methods | Point anomaly | Quantitative measures | UCI datasets [51] |
| Cao *et al.* [3] | k-NN, OCSVM | Price manipulation | AUC/ROC curve | NASDAQ |
| Li *et al.* [8] | k-NN, DT, LR, SVM, ANN | Trade-based manipulation | AUC | China Security Regulatory Commission (CSRC) |

stock market trading [49]. Note that such techniques imply that normal instances are far more frequent than anomalies in the test data. However, if this assumption is not true, such techniques may suffer from many false positives.

To reduce the number of false positives, Das *et al.* [40] proposed an Active Anomaly Detection (AAD) technique that incorporates expert (human) feedback into an ensemble of anomaly detectors. The AAD method tries to maximize the number of true anomalies presented to the expert analyst. Anomalies are internally ranked in every interactive feedback loop and presented to the expert who assigns a true label, either anomalous or nominal. A tree-based anomaly detector can be treated as ensembles to incorporate the feedback into them by employing AAD, such as Isolation Forest [56].

While considering machine learning-based approaches for detecting manipulations in financial markets, it is crucial to understand the following. First, there is scarcity of labeled data. Second, data is highly imbalanced. Third, manipulative market behaviors generally span over a sequence of anomalous events. Supervised anomaly detection methods, such as k-NN, RT, ANN, and SVM, are applied for financial market manipulations detection when the fully-labeled dataset is available [3], [6], [8], [32]. Supervised anomaly detection methods, including [32], [49], [6], and [3] provide generic approaches for financial domain monitoring because these can be used to detect adversarial market behaviors, such as trade-based manipulation where one or more actors work in collaboration, and they often have domain knowledge. These methods perform well in comparison to the traditional statistical methods when the notion of anomaly is clearly defined [3]. Misclassification of instances of minor/anomalous class is another common problem with the supervised anomaly detection that occurs due to the class-imbalance problem [30]. Although there exist several solutions [34], [57], [58] that combine machine learning algorithms with sampling methods to overcome this problem, the generated list of false-positive alerts has equal weights in terms of severity. There exist learning-to-rank method [59] that produce sorted lists of alerts, ranked in terms of their severity. Frery *et al.* [48] provide a recent learning-to-rank method for highly imbalanced data based on an average precision approach.

Unsupervised anomaly detection methods are the default choice [41] for unlabeled data. Ahmed *et al.* [13] and Ahmed *et al.* [49] used unsupervised methods to detect point anomalies in financial market data. This is a promising approach for financial market surveillance due to the following. First, it can be easily generalized for different application domains where data is continuously evolving. Second, since it considers only the internal structure of the dataset to detect anomalies, different notions of anomalies, e.g., contextual and collective, can be defined for market monitoring by domain experts. The nearest neighbor based unsupervised methods, such as Local Outlier Factor, Connectivity-based Outlier Factor, and Local Outlier Probability, perform better if the task is to detect contextual/local anomalies. However, these methods are prone to generating a large number of false-positives if applied for detecting point/global anomalies. On the other hand, although the k-NN is more suitable for point/global anomalies, its performance in terms of generated false-positives is average in the case of contextual/local anomalies. Therefore, the k-NN should be preferred when the nature (point vs. contextual or collective) of anomalies is not well defined. A large number of false-positives generated by machine learning-based approaches used for detection of financial market manipulations makes it challenging for human experts to review/analyze all alerts. Active Anomaly Detection [40], [50] is a human-in-loop learning method in which the designed framework interacts with the expert or the information sources to assign true labels. This approach has the potential for manipulation detection in the financial

domain due to its ability to substantially reduce the number of false-positives.

### B. NATURE OF INPUT DATA

Time-series data is a sequence of data instances taken successively at equal intervals of time. ARIMA is a general-purpose technique for modeling temporal data with seasonality [60]. Although it effectively detects anomalies in data with regular hourly, daily or weekly patterns, it fails in dynamically determining the period of seasonality.

Contrary to traditional anomaly detection, the time series anomaly detection is not well-understood, and time series anomaly detection techniques proposed over time have been spread across several application domains, including detection of anomalous heartbeat pulses in ECG data [61], cyber-attack detection in recommender systems [62], and detection of flight anomalies using sensor data from aircraft [63]. In the following, we review time series anomaly detection methods used in different application areas and discuss issues involved in their design.

#### 1) TIME SERIES ANOMALY DETECTION

In time-series data, there are many ways in which anomalies may occur. First, anomalies can be the individual data instances that vary significantly with respect to other data instances in the time series. Second, a subsequence within the time series can be anomalous with respect to long sequences. Third, the entire time series can be anomalous with respect to the time-series database [28].

There are important machine learning applications where data represents a sequence of events, and each event occurs at a given point in time. Anomaly detection in time series typically involves identifying subsequences within a time series that mismatches significantly with respect to the remaining time series. Note that the entire time series could also be treated as an anomaly. Time series anomaly detection requires extracting windows from the time series and then applying machine learning methods to detect anomalous subsequences. The subsequences are first transformed into either vector space or discrete space and then compared to detect anomalies. In a vector space representation, each subsequence can be represented by a multi-dimensional vector, and therefore traditional proximity-based machine learning methods can be used for anomaly detection. When both the testing and training time series sequences are of equal lengths, a simple Euclidean distance measure can be used to compute the proximity. However, when the testing and training time series sequences are of different lengths, such a simple distance measure does not work due to its inability to capture existing feature correlations. A more complex measure, such as Dynamic Time Wrapping (DTW) [64] is more suitable for comparing time series of different lengths.

In time series anomaly detection, a prediction-based approach is commonly used, where machine learning-based regression models are used to forecast future time series by using historical time-series data. In this approach, if the predicted time series deviates significantly from the actual time series, the system will consider it an anomaly. In this approach, any regression-based model can be used for the prediction of time series. For example, Extensible Generic Anomaly Detection System (EGADS) uses a set of default models such as ARIMA [73], Exponential Smoothing [74] and Kalman Filter [75] to model and forecast the time series.

Golmohammadi and Zaiane [67] proposed a prediction-based Contextual Anomaly Detection (CAD) method for time-series data of financial securities. The algorithm can identify contextual/local anomalies within the group of similar time series that do not follow seasonal patterns and are non-homogeneous. Instead of predicting the following values using historical time-series data, the CAD exploits the behavior of similar time series to predict expected values. A subset of a given time series is selected based on a given window size, and then the centroid of the time series is calculated that represents the expected behavior of the time series of the group within the window. The centroid can be calculated with the help of the mean or weighted mean of values within the window. The Pearson correlation coefficient [76] between each time series with the centroid is used to predict values of the time series, and in the end, an anomaly score is calculated by using Euclidean distance of the predicted value and the actual value of the given time series.

In prediction-based anomaly detection, a significant deviation between predicted and real values is identified as an anomaly. These methods produce alerts for each identified potential anomaly, and the alerts are further analyzed by human experts/analysts. Thus, the system must identify and remove alerts generated by spurious events so that the most relevant alerts are passed to human analysts. Laptev *et al.* [68] introduced a generic and scalable framework for automated anomaly detection on large time-series data that reduces the number of false positives.

In many applications, such as financial markets, abnormal behavior is simply not one single event but a series of events that occur in a sequence spanning over a time period. Therefore, detecting such abnormal behaviors requires capturing both the sequential order and the time window in which the anomalous set of events occurs. As listed in Table 2, Mannila *et al.* [65], Atallah *et al.* [11] and Rossi *et al.* [66] all suggested methods for abnormal event detection based on windowing and event sequencing techniques.

In the emerging area of Deep Learning, Shipmon *et al.* [31] presented a study on Google stream data to capture unexpected peaks and drops in network traffic. DNNs, RNNs and LSTM are some of the recent deep neural networks commonly used to detect anomalies by forecasting future values compared with actual values. Next, an error is calculated according to an anomaly detection rule, using, e.g., Gaussian tail probability, which is then used to generate an alert. Deep anomaly detection techniques learn hierarchical discriminative features from data. This automatic feature learning capability eliminates the need for engineering features manually by domain experts. It, therefore, has shown potential

**TABLE 2.** State-of-the-art machine learning methods and statistical techniques used for anomaly detection in sequential/time-series data.

| Paper | Methods | Manipulation type | Evaluation measures | Data |
|---|---|---|---|---|
| Mannila *et al.* [65] | Event sequencing using windowing technique | Trade-based manipulation | Sensitivity, F2 measure, Specificity | – |
| Atallah *et al.* [11] | Event sequencing using windowing technique | Collective, Contextual | Qualitative measures | Walmart data |
| Rossi *et al.* [66] | Frequent item set mining using windowing technique and categorical clustering | Collective, Contextual | Sensitivity, Clustering silhouette | Smart meter data |
| Golmohammadi and Zaiane [67] | CAD | Collective, Contextual | F-measure, Precision, Recall | S&P500 index |
| Laptev *et al.* [68] | ARIMA, Kalman filtering | Collective, Contextual | F1-Score | Time-series data |
| Shipmon *et al.* [31] | DNN, RNN, LSTM | Collective, Contextual | Confusion matrices, Recall, Precision | Google stream data |
| Zhu and Laptev [69] | Bayesian deep model | Collective, Contextual | – | Uber cab data |
| Munir *et al.* [70] | CNN | Point, Contextual | F-score | Yahoo Webscope |
| Zhang *et al.* [71] | Variational autoencoder | Collective, Contextual | AUC | UCR [72], UCI |

to solve the end-to-end problem, taking raw input data in the involved domains, such as text and speech recognition. Munir *et al.* [70] proposed an unsupervised deep anomaly detection technique capable of detecting a wide range of anomalies such as point, contextual, and collective anomalies in time-series data. It uses unlabeled data to capture the data distribution used to forecast the normal behavior of a time series. A similar study by Zhang *et al.* [71] proposed a time series anomaly detection method based Variational AutoEncoder model (VAE) with re-Encoder and Latent Constraint network (VELC). The authors emphasized the importance of accurately predicting time series and reliably estimating prediction uncertainty for anomaly detection. Although this problem is challenging, especially during high variance segments such as holiday and sports events, probabilistic time series forecasting can make predictions of such high variance data. In [69], an end-to-end Bayesian deep model was proposed that gives time series forecasting with uncertainty estimation, which could be used for large-scale anomaly detection. A detailed survey on deep anomaly detection is presented by [16], [17].

For financial market monitoring, the temporal aspect of data provides crucial information for detecting abnormal market behaviors. Different types of anomalies (point, contextual or collective) can be defined in time-series data, and then anomalies are detected using a prediction-based approach. Simple anomalous behaviors can be represented by data instances that deviate significantly in a given time series. These anomalies, which are generally contextual in nature, are detected using the prediction-based approach, where both traditional statistical models [73] and machine learning methods [67] can be applied. More complex anomalous behaviors, such as adversarial market behaviors, are often represented by a sequence of events that spread over time-series, and therefore detecting them requires identifying unusual shape subsequences in the time series. Traditional distance-based or density-based anomaly detection techniques cannot detect seasonal or periodic anomalies that are often seen in time-series data. To detect such behaviors, subsequences are transformed to other forms, such as multidimensional vectors or discrete sequences, and then windowing and sequencing [11], [65] methods are applied to detect abnormal events.

A market manipulation detection system generally consists of two components [68]: a prediction module and an anomaly detection module. As mentioned above, the prediction module predicts future values of time series and the anomaly detection module identifies the instances that deviate significantly. These anomalies indicate abnormal market behavior, and are used for generating alerts in a market monitoring system. However, this is a challenge that noise in data often produces a large number of false positives, making it difficult for human analysts to process all the alerts and detect real cases of manipulations. Recently, Deep Learning methods have performed exceedingly well in a variety of machine learning tasks. An extraordinary feature of Deep Learning methods is their ability to learn from the characteristics of data by transforming it into higher dimensions, and therefore these methods can serve as an alternative to the traditional machine learning methods. Recently, several Deep Learning

methods, such as [70], [31] and [71] has been used for financial market anomaly detection tasks. In general, the deep learning techniques require a large amount of data to generate accurate results; however, some of the techniques, such as [70] can be trained using a relatively small dataset. Despite being scalable and robust, these methods fall in the 'black-box' category.

### C. SUMMARY

In this section, we summarize the main points of the literature reviewed above. The machine learning methods underlying financial market surveillance systems and other related applications can be categorized broadly into two main categories: 1) availability of labeled data and 2) the nature of input data. The first category covers SAD methods that have been used for anomaly detection tasks where data have plenty of correctly classified instances of anomalies, and there is no time component. The SAD methods, such as RF, k-NN, DT, ANN, and SVM, perform well in balanced datasets. However, recall that anomalies are inherently very few, and therefore even if there are labels in a dataset, it is generally difficult for the SAD methods to classify the anomalies correctly. This means that in the case of highly imbalanced data, SAD methods often give high accuracy due to correctly categorizing the majority class but often misclassifying the minority class in the data. Therefore, the imbalanced data often needs some preprocessing by either oversampling the minority class or undersampling the majority class so that a supervised model can learn to classify both patterns correctly.

Since labeled data is rare and costly to produce, the SSAD and UAD techniques such as OCSVM, iForest, clustering methods, and LOF are used in anomaly detection tasks. These methods mainly work on the distance measures that define a boundary between normal and abnormal data instances. However, since separating the two types of instances is a non-trivial task due to the absence of clearly defined boundaries, it becomes challenging to evaluate the performance of these weakly supervised machine learning methods in anomaly detection tasks.

In the second category, which is more relevant to ML-based financial market surveillance, we reviewed papers on anomaly detection in time-series data. In time-series data, anomalies can occur in many ways, such as a point could deviate significantly from the rest of the series, or a subseries could be anomalous. A variety of methods, ranging from the statistical methods [60], [77], e.g., ARMA and ARIMA, and machine learning methods [67] to recent deep learning methods [16], have been used for anomaly detection in time-series data. In addition, event sequencing methods with windowing have also been proposed to detect collective and contextual anomalies. One of the important factors in the time series anomaly detection is the stationarity of time series. If a time series is not stationary, a model describing it will vary in accuracy at different time points. Its time components, e.g., trend and seasonality, should be removed during preprocessing to make a time series stationary.

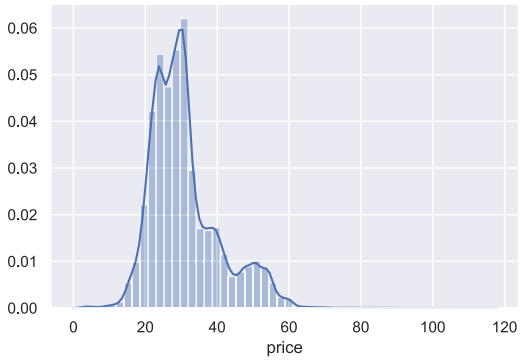



**FIGURE 4.** Spot price distribution of price area NO2.

## IV. EXPERIMENTS

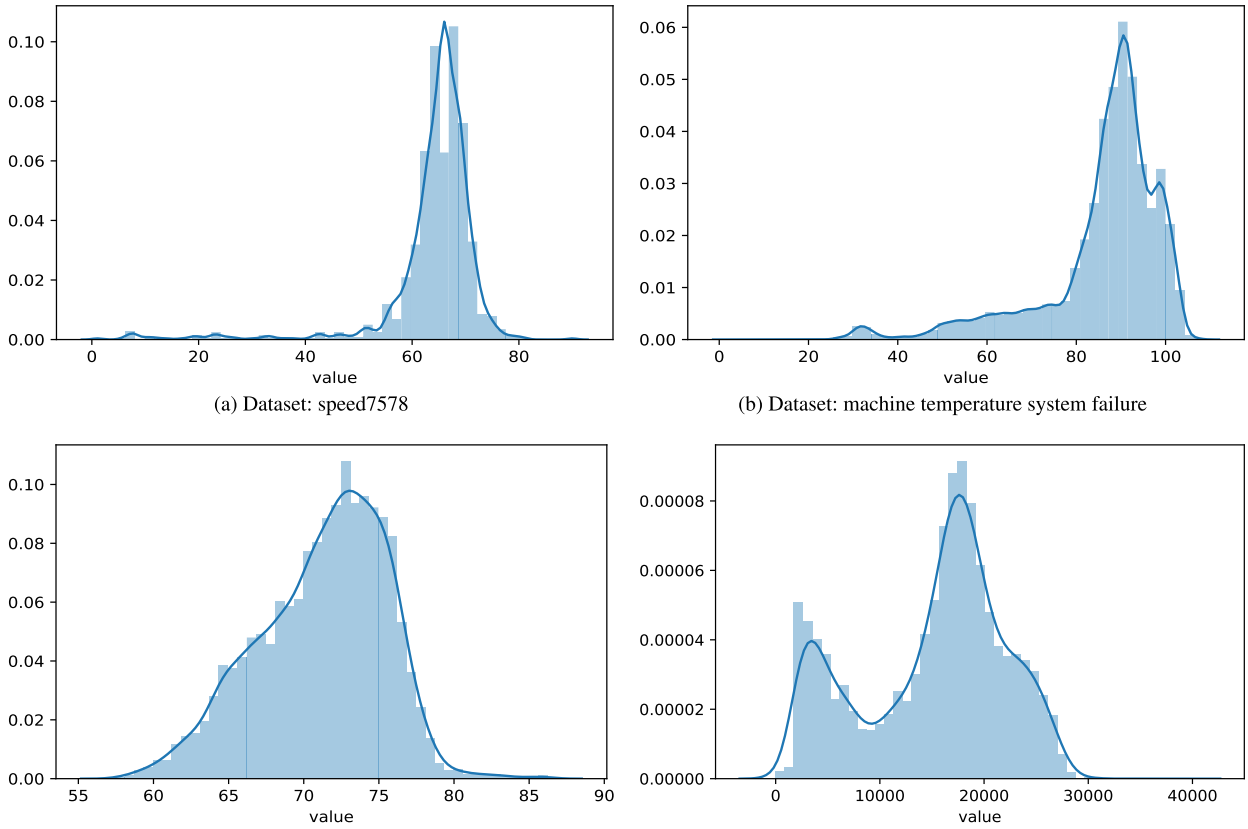
In this section, we systematically evaluate and compare some of the anomaly detection methods which we reviewed in Section III. We perform our experiments on two different time-series datasets: Elspot and NAB. We apply both supervised and unsupervised methods to detect price anomalies in Elspot data. The main motivation to include a comparative experimental study in this paper is to give the reader an even deeper insight into choosing between different machine learning methods for anomaly detection in financial time-series data when labels are absent.

### A. DATA

In this paper, we experimented with different machine learning methods for detecting anomalies in two time-series datasets. The first dataset that we used for our experiment is taken from the NordPool's Elspot market. As mentioned in Section II-A, Elspot is a day-ahead auction market where power contracts are traded for next-day physical power delivery. The Elspot trading market is divided into 24 price areas, each representing one bidding area or a constellation with a common price. Figure 4 shows the distribution of electricity prices of one such price area, 'NO2'.

We collected NordPool's Elspot data from January 2016 to September 2018, which contains hourly spot prices, the volume of electricity (demand and production), price area, and timestamp. As discussed further in this section, we used different supervised and unsupervised ML methods to detect price anomalies in this dataset. In Section V-A2 we forecast the next day's electricity prices for each hour and then classify the prices as anomalous or normal based on their values. If we want the forecasting model to work well, it is also essential to convert non-stationary time series to a stationary time series. Because most of the time-series forecasting methods are based on statistics, the assumption is that the training and test sets are drawn from the same distribution. Therefore, the methods may not perform well if the future (test) data distribution is different from the past (training) data.

The other data we used is the Numenta Anomaly Benchmark (NAB) dataset which provides a benchmark for evaluating anomaly detection algorithms in high-velocity online applications. It comprises over 50 labeled real-world and




(a) Dataset: speed7578

(b) Dataset: machine temperature system failure

(c) Dataset: ambient temperature system failure

(d) Dataset: nyc taxi

**FIGURE 5.** Distributions of NAB Datasets, where "value" represents: (a) speed, (b) temperature, (c) temperature (d) total number of taxi passengers in every 30 minute bucket.

artificial time-series data files, together with a novel scoring mechanism designed for real-time applications. Each row in a NAB's time-series data contains a timestamp plus a single scalar value. Anomaly detection in the NAB's time-series data has practical and significant applications across many industries such as preventative maintenance, fraud prevention, fault detection and monitoring in finance, IT, security, e-commerce. Anomalous patterns differ significantly across application domains, e.g., a latency of one second in periodic ECG data could be a significant fluctuation, but the same pattern in electricity trading may be completely normal. Thus, it is important to include time series from a variety of domains and applications. We chose datasets with known causes of anomalies, such as ''ambient temperature system failure2'', ''nyc taxi'' and ''machine temperature system failure''. The others are real-time traffic data ''speed7578'', including speed from specific sensors, and a real tweets' data. The real tweets dataset is a collection of Twitter mentions of large publicly traded companies such as Google. Figure 5 shows distributions of these time-series data.

We cannot quantitatively evaluate the methods and compare them since we do not have any labels in our Elspot dataset. We can only qualitatively analyze the methods by looking at the results from different methods or including expert's feedback. Therefore, we decided to perform experiments on NAB dataset. As mentioned earlier, both datasets (Elspot and NAB) are time-series datasets. Therefore, the underlying properties such as trend and seasonality are common in both datasets.

### B. METHODOLOGY

In machine learning-based predictive analysis, models are first trained on historical data and then used to predict future observations. Predictive analysis is a branch of advanced analytics used to make predictions about unknown future events using machine learning, data mining, statistics, and artificial intelligence. Such analysis could help predict future anomalous events. Suppose input data does not have instances outside the labeled categories, and there are no complicated trends or patterns, a supervised machine learning method could be used to detect anomalies in such data.

On the other hand, unsupervised machine learning methods first learn characteristics of unlabeled data during training. The data instances which deviate too much from the normal are considered anomalous. In principle, unsupervised methods can be used to develop an anomaly detection system that can detect any type of anomalies, including ones that have never been seen before. However, deciding what is anomalous is a significant challenge. For example, in the wholesale power market, spikes in electricity prices are expected from one hour to the other in some circumstances.

In contrast, it may also be the case that high price results from changes in supplier behavior itself to exploit particular market circumstances.

In the following, we discuss different supervised and unsupervised machine learning methods that we use in our analysis to detect anomalies in the power market time-series datasets.

- **Quantile Regression Forest (QRF):**
  QRF is proposed in [78] which is a generalization of RF [79]. Random Forest (RF) provides an accurate approximation of the conditional mean of a response variable, and it also includes information about the full conditional distribution of the response variable. QRF, which can infer conditional quantiles, provides a non-parametric and accurate way to estimate conditional quantiles for high-dimensional predictor variables. Let $Y$ be a true response variable and $X$ possibly a high-dimensional predictor variable. The goal of standard statistical analysis is to infer the relationship between $Y$ and $X$. Standard regression estimates the conditional mean of the response variable $Y$, given $X = x$, whereas quantiles provide complete information about the distribution of $Y$ as a function of the predictor variable $X$. The prediction then returns the mean and full conditional distribution $P(Y \leq y|X = x)$ of response values for every x. Using the distribution, it is trivial to create prediction intervals for new instances simply by using the appropriate percentiles of the distribution.
- **Gradient Boosting Regressor (GBR):**
  In machine learning, ''Boosting'' combines multiple simple models into a single composite model [80], also referred to as ensemble or additive model. In the process, weak learners (simple models, e.g., decision trees) are added one at a time while keeping existing trees in the model unchanged. By adding more and more weak learners, the final complete model becomes a strong learner. Decision trees are used as a weak learner in gradient boosting, and specifically, regression trees are used that output real values for splits and whose output can be added together—allowing subsequent models output to be added and ''correct'' the residuals in the prediction. The residual in GBR is the difference between the current prediction and the known correct target value.
- **Extra Tree Regressor (ETR):**
  Extremely randomized trees (or extra-trees (ET)) algorithm [81] is a relatively recent approach that shares several characteristics with RF, and taking the randomness in the tree splits a step further. Similar to RF, ET uses a random subset of features to train each base estimator. However, instead of choosing the most discriminating split in each node, the algorithm picks the best among $k$ randomly generated splits. Another difference between RF and ET is that the latter uses the whole training data set to train each regression tree instead of a bootstrap sample in RF. The rationale behind ET is that the explicit randomization of the split points is expected to reduce variance more than other methods with weaker randomization schemes. Using the complete training data, rather than a sample of them, is motivated by reducing the model's bias.
- **k-nearest Neighbors Global Anomaly Detection:**
  The k-NN method focuses on detecting global anomalies instead of local ones. Each data point looks for k nearest neighbors in a dataset and then computes the anomaly score using either the distance to $k^{th}$ nearest-neighbor [38] or average distance [82] to all k-nearest neighbors. However, the value of the absolute score depends on the dataset itself, the number of dimensions, and normalization.
- **Local Outlier Factor:**
  Breunig *et al.* [83] proposed an unsupervised anomaly detection algorithm termed Local Outlier Factor. The LOF algorithm uses a concept of local density, which is measured in terms of the typical distance decided by k-nearest neighbors of a given data point, to detect anomalous data points. The LOF of a data point gives its point density compared to the density of its neighbors. Therefore, if the density of a point is significantly smaller than the densities of its neighbors ($LOF \gg 1$), the point is an outlier. The LOF algorithm is a useful unsupervised anomaly detection method for situations where anomalous and normal points do not clearly define boundaries.
- **One-class Support Vector Machine:**
  One-class support vector machines [39], [84] are often used for semi-supervised anomaly detection, where OCSVM is trained only on normal data; later, it classifies anomalies and normal data in the test set. To identify anomalous observations, an OCSVM estimates a distribution that encompasses most of the observations and labels them as anomalous ones that lie far from it concerning a suitable metric. Although OCSVM is heavily used as a semi-supervised anomaly detection method, it uses a soft margin and is an unsupervised algorithm by design. In the unsupervised anomaly detection scenario, the OCSVM is trained using the dataset, and afterwards, each instance in the dataset is scored by a normalized distance to the determined decision boundary.
- **Isolation Forest:**
  Isolation Forest [37] is different from other outlier detection methods. Instead of profiling normal data points, it explicitly identifies anomalies. It is similar to any tree ensemble method and built on top of decision trees. The partitions are created by randomly selecting the features and then selecting a random split between the minimum and maximum value of the selected feature; thus, a tree has been made. In principle, outliers are less frequent than normal observations, and they lie further away from the normal observations in the feature space. Therefore, outliers should be identified closer to the tree's root with fewer splits necessary using such random partitioning.

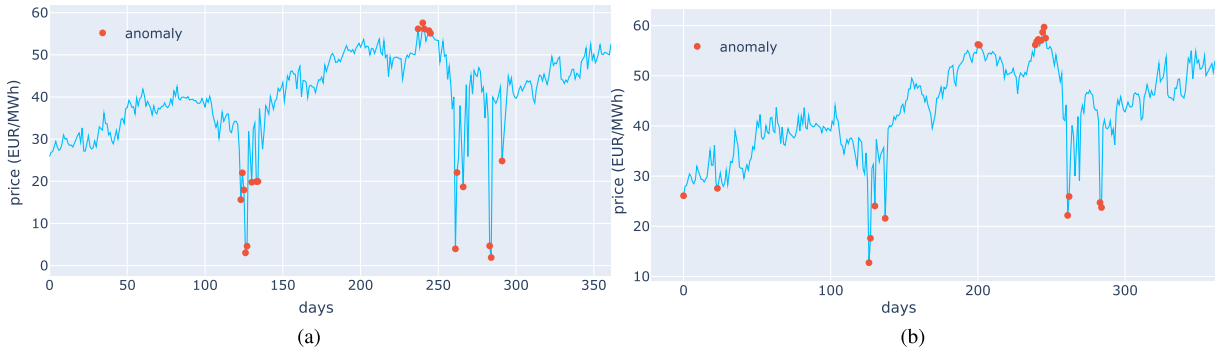



**FIGURE 6.** Anomalies detected using iForest in ELSPOT data from 2018-01-01 to 2018-09-30. (a) anomalies in "hour 01" of spot prices, and (b) anomalies in "hour 03" of spot prices.

Above all, popular ML methods such as QRF, GBR, ETR, k-NN, kmeans, LOF, OCSVM, and iForest were described. LOF and iForest have been widely used in outlier detection in various domains. iForest is sensitive for global outliers, whereas LOF performs well in local outlier detection but has high time complexity. LOF is useful in capturing outliers in the power market when some hours' electricity price deviates much from the local density. The primary motivation of using iForest in the analysis is, iForest explicitly identifies anomalies instead of profiling normal data points. We do not have any labels in our power market data; hence, iForest could be suitable for the analysis. Also, outliers are rare in the data; thus, the class imbalance problem arises. When the ratio between classes in the data is, e.g., 1:100 or larger, an early attempt to model the problem will give very high accuracy but very low specificity. From that perspective, OC-SVM could be a good candidate algorithm for analyzing NAB and power market data. Decision Trees are easy to interpret and explain. They are typically fast and scalable; however, they are prone to overfitting, but ensemble methods such as random forest and boosted trees can overcome this issue. Since we do not have labels in the power market dataset, QRF, GBR, and ETR are good candidates to make future electricity price predictions with prediction intervals, and the prices far beyond the interval are considered outliers.

## V. RESULTS AND DISCUSSION

In the following, we analyze and discuss anomaly detection results obtained by using different machine learning methods to the datasets introduced in Section IV-A.

### A. RESULTS

#### 1) ELSPOT PRICE DATASET

We experimented with the NordPool's Elspot dataset, mentioned earlier in Section IV-A. Elspot is a day-ahead electricity trading market, where all the bids and offers are submitted, and then the market-clearing prices for each hour of the next day are determined. We used spot price data from January 2018 to December 2018 in our analysis. Since this is time-series data, it has components, such as seasonality and trend. It is essential to remove these time-varying components from the time series so that machine learning methods can learn patterns in data. We also performed experiments on longer time series with absolute price values from 2016 to 2018, where it was difficult for machine learning methods to learn the patterns in the raw dataset. However, the models performed well when we trained them with the lagged features (*PriceDelta*) along with absolute price values. To remove trend and seasonality components from the dataset, we calculated price difference using Equation 1:

$$PriceDelta_{d_{h1}} = Price_{d_{h1}} - Price_{(d-1)_{h1}} \tag{1}$$

where $Price_{d_{h1}}$ is the current day's hour 1 price and $Price_{(d-1)_{h1}}$ is the previous day's hour 1 price.

We also arranged data in a way that all the delta prices of day $d1$ will be in a single row. First, we performed experiments on the preprocessed dataset using unsupervised machine learning methods. Figure 6 shows the daily prices of an hour generated from such an experiment using the iForest method. Figure 6(a) shows the daily prices of hour 1 in the dataset, where red color dots highlight outliers. Similarly, Figure 6(b) shows the daily prices of hour 3 in the dataset. Figure 7 shows outlying days detected using different unsupervised ML methods in the dataset. The iForest method learns to define a clear separation between normal and outlier days. These analyses and visual representations demonstrate that unsupervised machine learning methods can capture normal behavior.

Figure 8 shows which methods agree with a detected anomaly. For example, in our data, "day 63" is predicted as an anomaly by kmeans, iForest, OCSVM, and k-NN but not by LOF. Therefore, the prediction for this day has the majority vote of being anomalous, whereas "day 236" is flagged as an anomaly by LOF only. Thus, it is specified as a normal point by majority vote. Since we do not have any labels in this dataset, it is hard to analyze the performance of these models quantitatively. However, the above voting method can be used to decide which points are more likely to be actual anomalies, and those points can be analyzed further. This approach would also reduce the number of alerts passed to human analysts, for example, in the case of unusual electricity prices in the Elspot data.

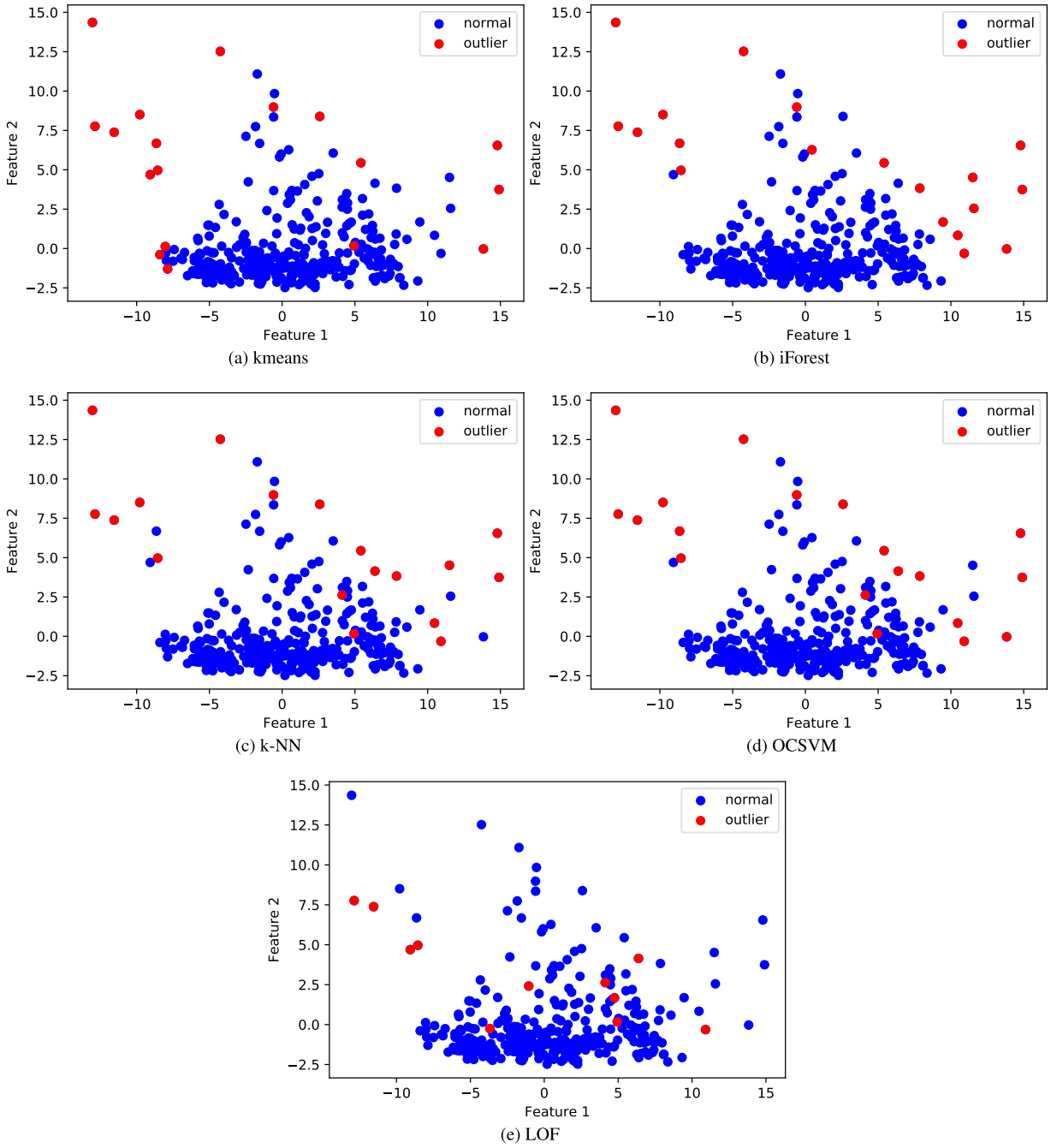



**FIGURE 7.** Outlier detection in Elspot price data using kmeans (a), iForest (b), k-NN (c), OCSVM (d), and LOF (e). Outlier days are marked with red dots.

### 2) PREDICTIVE ANALYSIS ON ELSPOT DATA

In our next set of experiments, we took a predictive analytical approach to identify anomalies in the Elspot dataset. Predictive analysis is about analyzing current and historical data to forecast the probability of future outcomes, i.e., to approximate a mapping function $f$ from input variables $X$ to a continuous output variable $Y$, which is a real value such as the electricity spot prices in this case. We perform predictive analysis on historical spot price data and predict future electricity spot prices using supervised machine learning methods. The data is preprocessed in the same way as described in Section V-A1. We also generate future price values ($y_f$) by shifting the sequence of price values with a time interval of t = 24h. We are interested in generating not only point prediction but also the prediction interval, where prediction interval is an estimate to an interval into which the future observation will fall within a given probability. For predictive analyses on the Elspot dataset, we applied tree-based regression methods: QRF, GBR, and ETR.

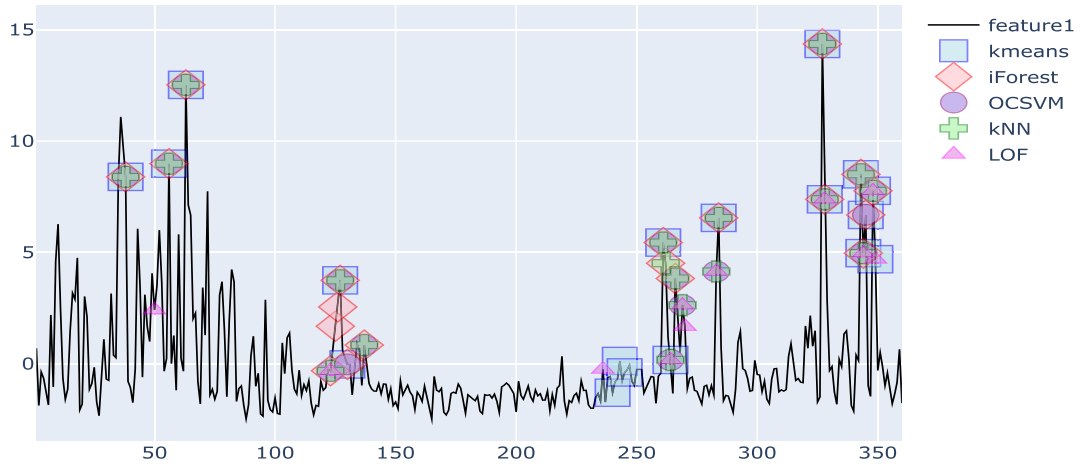



**FIGURE 8.** A plot showing identified outlying days on ELSPOT data using different unsupervised learning methods.

**TABLE 3.** Comparison of different tree-based regressors on Elspot data.

| Method | MSE | MAE | RMSE | Score |
|---|---|---|---|---|
| QRF | 29.85 | 2.96 | 5.4 | 0.51 |
| GBR | 29.07 | 2.94 | 5.3 | 0.51 |
| ETR | 30.0 | 2.96 | 5.47 | 0.51 |

We divide the dataset into train and test sets, then train models using QRF, GBR, and ETR on the training set and make predictions on test data to evaluate the model performance. The model performance is evaluated by using standard metrics, such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the score $R^2$ as defined in equation 2, which is a coefficient of prediction determination. The best model can have a score of 1.0, while the worst model can get a negative score. Table 3 provides evaluations of the three models trained and tested on Elspot data. Since our aim in this work is to demonstrate possible machine learning methods to identify anomalies, we trained these models only with the default parameters

$$R^2 = (1 - u/v) \tag{2}$$

The residual sum of squares ($u$) and total sum of squares ($v$) can be found using the formulas below:

$$u = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2) \tag{3}$$

$$v = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2) \tag{4}$$

where $y_i$ is the observed value, $\hat{y}_i$ is the value estimated by the regression line, and $\bar{y}_i$ is the mean value of a sample.

Once we get the day-ahead price prediction, we can use these predicted values and prediction interval to identify anomalies in the test set. Figure 9 shows the point predictions and prediction intervals generated by QRF. The true values lying outside the prediction interval are considered as outliers. Additionally, we rank the outliers based on their deviations from the mean prediction i.e., points with large deviations are highly anomalous, whereas points close to the prediction interval are considered as low anomalous points.

#### 3) NAB DATASET

As we mentioned in section IV-A, due to the lack of labels in Elspot data, we cannot quantitatively evaluate the machine learning-based anomaly detection models on this data. However, NAB time-series datasets are fully labeled, and therefore we can easily evaluate the machine learning-based anomaly detection methods by using standard evaluation measures mentioned in Section II-C. Table 4 presents the performance of different unsupervised machine learning methods applied to NAB datasets. High accuracy in all cases can be attributed to the highly imbalanced nature of NAB datasets. We have datasets where true positive cases are very rare; hence model performs well in predicting points from the majority class. We also observe that some of the unsupervised methods are capable of identifying anomalies with good TPR in the speed7578 dataset (Table 4a) and ambient temperature system failure dataset (Table 4b).

Table 5 presents the performance of different supervised machine learning methods applied to NAB datasets. In supervised machine learning methods, we sampled training set from the minority class and tested three different ML models on the test set that contains only a few outliers. Logistic regression (LR) captured some outliers, but the Random Forest classifier (RFC) and k-NN could not classify outliers correctly in the test sets.

### B. DISCUSSION

The main objective with the experiments is to highlight the behaviour of the discussed methods by analyzing financial market data, i.e. our Elstpot dataset. Our first experiment predicted abnormal price days using different unsupervised machine learning on our Elspot dataset. We could not evaluate the performance of these methods using standard evaluation metrics due to absence of labels in the data. Instead, we used a concept of majority votes to roughly agree on days predicted

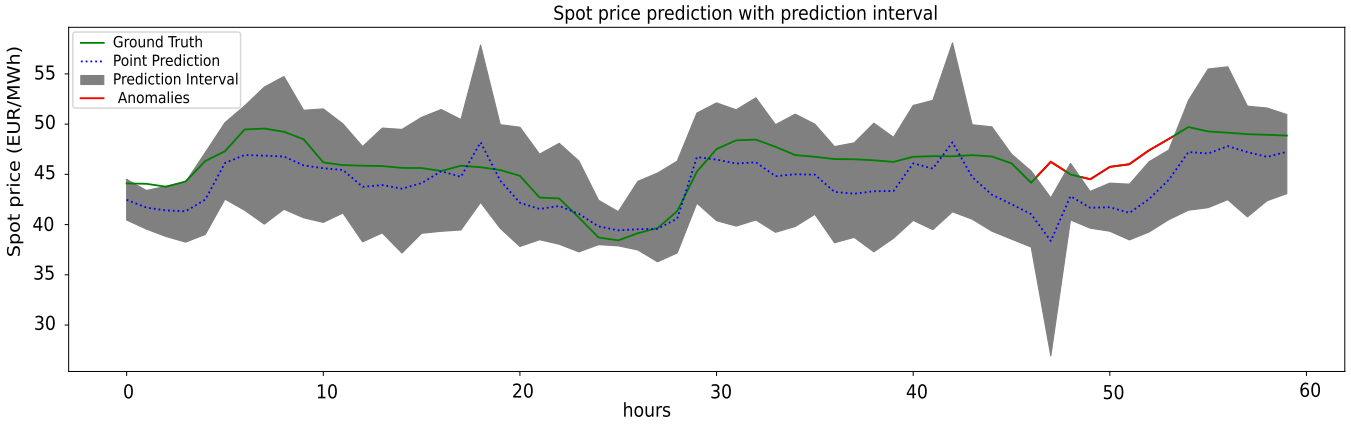



**FIGURE 9.** Point prediction and prediction interval generated by QRF.

**TABLE 4.** Comparison of different unsupervised machine learning methods to detect anomalies in NAB dataset.

(a) speed7578

| Method | TPR | TNR | FPR | AUC |
|---|---|---|---|---|
| kmeans | 0.33 | 1.0 | 0 | 0.66 |
| iForest | 0.66 | 0.95 | 0.04 | 0.8 |
| OCSVM | 0 | 0.99 | 0.002 | 0.49 |
| k-NN | 0 | 0.99 | 0.0008 | 0 |
| LOF | 0.66 | 0.97 | 0.02 | 0.81 |

(b) ambient temperature system failure

| Method | TPR | TNR | FPR | AUC |
|---|---|---|---|---|
| kmeans | 0.5 | 0.99 | 0.008 | 0.74 |
| iForest | 1.0 | 0.95 | 0.04 | 0.97 |
| OCSVM | 0.5 | 0.99 | 0.0009 | 0.74 |
| k-NN | 0 | 0.99 | 0.0009 | 0.49 |
| LOF | 0 | 0.97 | 0.024 | 0.48 |

(c) nyc taxi

| Method | TPR | TNR | FPR | AUC |
|---|---|---|---|---|
| kmeans | 0 | 0.99 | 0.0 | 0.49 |
| iForest | 0.2 | 0.95 | 0.049 | 0.57 |
| OCSVM | 0 .2 | 0.99 | 0.0009 | 0.59 |
| k-NN | 0.2 | 0.99 | 0.0004 | 0.6 |
| LOF | 0 | 0.97 | 0.025 | 0.48 |

(d) machine temperature system failure

| Method | TPR | TNR | FPR | AUC |
|---|---|---|---|---|
| kmeans | 0.25 | 0.99 | 0.0009 | 0.62 |
| iForest | 0.5 | 0.95 | 0.04 | 0.6 |
| OCSVM | 0.25 | 0.99 | 0.0009 | 0.62 |
| k-NN | 0.25 | 0.99 | 0.0004 | 0.62 |
| LOF | 0.25 | 0.97 | 0.02 | 0.61 |

**TABLE 5.** Comparison of different supervised machine learning methods on NAB dataset where *ambient temp* is ambient temperature system failure dataset and *machine temp* is machine temperature system failure dataset.

| Data | Method | TPR | TNR | FPR |
|---|---|---|---|---|
| speed7578 | RFC | 0 | 1.0 | 0 |
| | LR | 0.5 | 0.94 | 0.06 |
| | k-NN | 0 | 1.0 | 0 |
| nyc taxi | RFC | 0 | 1.0 | 0 |
| | LR | 0.33 | 0.68 | 0.31 |
| | k-NN | 0 | 1.0 | 0 |
| ambient temp | RFC | 0.0 | 1.0 | 0.0 |
| | LR | 0.0 | 1.0 | 0.0 |
| | k-NN | 0 | 1.0 | 0.0 |
| machine temp | RFC | 0 | 1.0 | 0 |
| | LR | 1.0 | 0.65 | 0.34 |
| | k-NN | 0 | 1.0 | 0 |

as 'abnormal' and use this majority vote to generate alerts. In the majority vote criteria, if a data point is predicted as an anomaly by four out of five methods, there is a good chance that the point is truly an anomaly. Although we cannot evaluate the final outcome here either, the above voting method can be used to decide which points are more likely to be actual anomalies, and those points can be passed to analysts for further investigation. Nonetheless, the method reduces the number of generated alerts to be passed on to human analysts, and therefore it can be a potential approach for machine learning-based automatic surveillance of financial markets.

Because unlabeled Elspot data prevents us from quantitatively comparing the methods we have discussed, and we can only analyse them qualitatively when using this dataset, we decided to perform additional experiments on the well-known and labelled NAB dataset. We compared different machine learning methods on the NAB dataset. Since we have labels in the NAB datasets, it is easier to evaluate the methods using standard evaluation metrics. Our experiments show that the unsupervised and semi-supervised learning methods performed better than the supervised learning methods on the NAB dataset. Although some of the time-series properties of both datasets (Elspot and NAB) are similar, it is not always guaranteed that the methods performing well on the NAB dataset would also perform equally well on the power market data or other financial market data. The anomaly detection problem is often domain-specific, therefore we cannot always generalize the detection techniques. Our experimental evaluation showed that the supervised methods such as k-NN perform well on labeled datasets, the unsupervised techniques such as clustering and LOF are suitable for unlabeled datasets, whereas semi-supervised techniques such as iForest and OCSVM perform better in anomaly detection tasks.

In our second experiment, we used a predictive analysis approach using ensemble methods (QRF, GBR, and ETR) to predict future electricity prices on the Elspot dataset. The ensemble methods performed well in predicting the mean

value of the response variable and the prediction interval, an estimation of the range of values (intervals) into which the future observations will fall with a given probability. The prediction intervals reveal the characteristics of underlying data and provide a simple way to sanity check the predictions. It is easy to generate alerts on a test set after we have a good trained model. Data points falling outside, e.g., 90% prediction intervals are treated as outliers and used for generating alerts. The severity of generated alerts is further assessed based on their deviations from the mean prediction. We evaluated our models using standard metrics such as MSE, MAE, and RMSE, but since we do not have any labels in the dataset, the quality of alerts could not be evaluated.

## VI. CONCLUSION AND FUTURE WORK

To develop a broad understanding of relevant approaches for implementing machine learning models for financial market surveillance, we first presented a broad review of general machine learning systems used in the financial markets (e.g. stock market surveillance) or transferable problems (abnormality detection, fraud detection, etc.). We then presented the example of physical power market to define what constitute an 'abnormal and suspicious market behavior' and how to explicitly map and formulate the regulator's expert knowledge and intuition in the domain, and use this when designing a machine learning based market monitoring system using electricity price data. Further, in order to highlight the behaviour of the discussed methods on financial data, we analyzed our Elspot dataset. Unfortunately, though, the Elspot dataset is, as so commonly seen in this area, unlabelled. This prevents us from quantitatively comparing the methods we have discussed, and we can only analyse them qualitatively when using this dataset. Therefore, we performed additional experiments on the well-known and labelled NAB data-set. Because both Elspot and NAB contain time-series data, some underlying properties (such as trend and seasonality) are common in both. While not ideal, we therefore believe that experiments on the NAB dataset can at least shed some additional light on the behaviour of the discussed methods, i.e., roughly indicate how the methods would fare on financial datasets based on their results on non-financial time-series data.

Financial market surveillance is a broad subject area due to a large variety of assets and nature of trading activities, making the market compliance space large and complex. Machine learning based approaches are emerging as promising technologies to fill this compliance space as solution providers seek new ways to evolve technology offerings to get ahead of compliance challenges. Altogether, the findings in this paper give information on designing machine learning based solutions for financial market surveillance by selecting the best models from a set of potential prediction algorithms. This approach would also make machine learning-based market surveillance systems explainable as each anomaly detection solution tested provides a root cause analysis for the anomalies detected, which could be used as feedback to the model for self-correction.

As future work, this approach can be improved by developing appropriate evaluation metrics for underlying machine learning methods [42], [43]. Another interesting direction of research in this approach is to further reduce false-positive alerts by incorporating feedbacks from analysts to analytic engine [40], [50]. The availability of benchmark datasets for such a study is very sparse. Hence, building a comprehensive synthetic dataset would be another interesting contribution to anomaly detection research. Interpretability/explainability of underlying machine learning models is an important issue in financial market surveillance. To mitigate the potential bias of humans and for transparent decision making, it is essential to understand how machine learning models make decisions [85]. In particular, since deep learning models are normally 'black-box' models [86], future research in this direction is needed to enable trust in the usage of these methods.

Deep learning methods have been shown to perform well in detecting point anomalies. These methods are known to learn characteristics of data by capturing complex dependency between features of data points. However, it remains to be explored in the future how this inherent strength of deep learning methods could be exploited to detect market manipulations represented by contextual or collective anomalies. Flexibility in building deep learning models by adding new features, such as neural layers, connections and objective functions, provides plenty of room to explore in this direction. Another future research direction is Deepfake [87], a synthetic content generation technology based on generative deep learning [88], that has recently made it possible to manipulate financial markets by digital impersonation [89], [90]. Detection of deepfakes is a technical challenge [87]. Due to the use of AI, social media and deep learning methods, detecting deepfake-based market manipulations is currently a challenge that should be addressed by deep learning research in the future.

## REFERENCES

[1] F. Black, "Toward a fully automated stock exchange, Part I," *Financial Analysts J.*, vol. 27, no. 4, pp. 28–35, Jul. 1971.
[2] F. Allen and D. Gale, "Stock-price manipulation," *Rev. Financial Stud.*, vol. 5, no. 3, pp. 503–529, Jul. 1992.
[3] Y. Cao, Y. Li, S. Coleman, A. Belatreche, and T. M. McGinnity, "Detecting price manipulation in the financial market," in *Proc. IEEE Conf. Comput. Intell. Financial Eng. Econ. (CIFEr)*, Mar. 2014, pp. 77–84.
[4] R. K. Aggarwal and G. Wu, "Stock market manipulations," *J. Bus.*, vol. 79, no. 4, pp. 1915–1953, Jul. 2006.
[5] E. J. Lee, K. S. Eom, and K. S. Park, "Microstructure-based manipulation: Strategic behavior and performance of spoofing traders," *J. Financial Markets*, vol. 16, no. 2, pp. 227–252, May 2013.
[6] H. Öğüt, M. M. Doğanay, and R. Aktaş, "Detecting stock-price manipulation in an emerging market: The case of Turkey," *Expert Syst. Appl.*, vol. 36, no. 9, pp. 11944–11949, Nov. 2009.
[7] D. Diaz, B. Theodoulidis, and P. Sampaio, "Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices," *Expert Syst. Appl.*, vol. 38, no. 10, pp. 12757–12771, Sep. 2011.
[8] A. Li, J. Wu, and Z. Liu, "Market manipulation detection based on classification methods," *Proc. Comput. Sci.*, vol. 122, pp. 788–795, Jan. 2017.

[9] M. Monster and R. J. V. BEng, "Heads or tails: Market surveillance and market abuse," Compact, The Netherlands, Tech. Rep., 2015, no. 4.
[10] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, p. 15, 2009.
[11] M. Atallah, R. Gwadera, and W. Szpankowski, "Detection of significant sets of episodes in event sequences," in *Proc. 4th IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2004, pp. 3–10.
[12] Nasdaq. (2017). *Smarts Market Surveillance*. [Online]. Available: http://www.nasdaq.com
[13] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Gener. Comput. Syst.*, vol. 55, pp. 278–288, Feb. 2016.
[14] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, 2004.
[15] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: A survey," *Data Mining Knowl. Discovery*, vol. 29, no. 3, pp. 626–688, 2015.
[16] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.
[17] G. Pang, C. Shen, L. Cao, and A. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, 2021.
[18] S. Ahmad, A. Lavin, S. Purdy, and Z. Agha, "Unsupervised real-time anomaly detection for streaming data," *Neurocomputing*, vol. 262, pp. 134–147, Nov. 2017.
[19] N. P. Spot, "The Nordic electricity exchange and the nordic model for a liberalized electricity market," *Nord Pool Spot, Norway*, 2009. [Online]. Available: http://www.nordpoolspot.com
[20] A. N. Campbell, "A guide to energy market manipulation," *Energy Law J.*, vol. 39, p. 177, 2018. [Online]. Available: http://www.eba-net.org/assets/1/6/Campbell_FINAL.pdf
[21] J. Örtenblad, "Market surveillance system," M.S. thesis, Dept. Financial Math., KTH, Stockholm, Sweden, 2001.
[22] S. Agrawal and J. Agrawal, "Survey on anomaly detection using data mining techniques," *Proc. Comput. Sci.*, vol. 60, pp. 708–713, Jan. 2015.
[23] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection," *Credit Scoring Credit Control VII*, vol. 7, pp. 235–255, Sep. 2001.
[24] S. Thiprungsri and M. Vasarhelyi, "Cluster analysis for anomaly detection in accounting data: An audit approach," *Int. J. Digit. Accounting Res.*, vol. 11, pp. 1–16, Jul. 2011.
[25] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," in *Proc. 16th Int. Conf. Pattern Recognit.*, vol. 4, Aug. 2002, pp. 385–388.
[26] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Comput. Secur.*, vol. 28, nos. 1–2, pp. 18–28, 2009.
[27] G. Ritter and M. T. Gallegos, "Outliers in statistical pattern recognition and an application to automatic chromosome classification," *Pattern Recognit. Lett.*, vol. 18, no. 6, pp. 525–539, Jun. 1997.
[28] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier Analysis*. Springer, 2017, pp. 1–34, doi: 10.1007/978-3-319-47578-3_1.
[29] X. Song, M. Wu, C. Jermaine, and S. Ranka, "Conditional anomaly detection," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 5, pp. 631–645, May 2007.
[30] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *Proc. 4th Int. Conf. Natural Comput.*, vol. 4, 2008, pp. 192–201.
[31] D. T. Shipmon, J. M. Gurevitch, P. M. Piselli, and S. T. Edwards, "Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data," 2017, *arXiv:1708.03665*.
[32] K. Golmohammadi, O. R. Zaiane, and D. Díaz, "Detecting stock market manipulation using supervised learning algorithms," in *Proc. Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2014, pp. 435–441.
[33] G. W. Schwert, "Anomalies and market efficiency," *Handbook Econ. Finance*, vol. 1, pp. 939–974, Jan. 2003.
[34] A. D. Pozzolo, O. Caelen, and G. Bontempi, "When is undersampling effective in unbalanced classification tasks?" in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2015, pp. 200–215, doi: 10.1007/978-3-319-23528-8_13.
[35] T. Marwala, *Computational Intelligence for Missing Data Imputation, Estimation and Management: Knowledge Optimization Techniques*. New York, NY, USA: Information Science Reference Hershey, 2009.
[36] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proc. Sci. Inf. Conf.*, Aug. 2014, pp. 372–378.
[37] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.
[38] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, 2000.
[39] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[40] S. Das, W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating expert feedback into active anomaly discovery," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 853–858.
[41] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, Apr. 2016, Art. no. e0152173.
[42] A. Lavin and S. Ahmad, "Evaluating real-time anomaly detection algorithms—The numenta anomaly benchmark," in *Proc. IEEE 14th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2015, pp. 38–44.
[43] N. Tatbul, T. J. Lee, S. Zdonik, M. Alam, and J. Gottschlich, "Precision and recall for time series," 2018, *arXiv:1803.03639*.
[44] C. Chatfield, "The holt-winters forecasting procedure," *J. Roy. Statist. Soc. C, Appl. Statist.*, vol. 27, no. 3, pp. 264–279, 1978.
[45] C. Chatfield, *The Analysis of Time Series: An Introduction*, 6th ed. Mar. 2016, pp. 1–329, doi: 10.4324/9780203491683.
[46] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen, "Sketch-based change detection: Methods, evaluation, and applications," in *Proc. ACM SIGCOMM Conf. Internet Meas. (IMC)*, 2003, pp. 234–247, doi: 10.1145/948205.948236.
[47] E. Jondeau, S.-H. Poon, and M. Rockinger, "Statistical properties of financial market data," in *Financial Modeling Under Non-Gaussian Distributions*. London, U.K.: Springer, 2007, pp. 7–32, doi: 10.1007/978-1-84628-696-4_2.
[48] J. Frery, A. Habrard, M. Sebban, O. Caelen, and L. He-Guelton, "Efficient top rank optimization with gradient boosting for supervised anomaly detection," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, 2017, pp. 20–35.
[49] M. Ahmed, N. Choudhury, and S. Uddin, "Anomaly detection on big data in financial markets," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 998–1001.
[50] S. Das, W.-K. Wong, A. Fern, T. G. Dietterich, and M. A. Siddiqui, "Incorporating feedback into tree-based anomaly detection," 2017, *arXiv:1708.09441*.
[51] A. Asuncion and D. Newman, "UCI machine learning repository," Irvine, CA, USA, Tech. Rep., 2007.
[52] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, and L. Tarassenko, "Static and dynamic novelty detection methods for jet engine health monitoring," *Phil. Trans. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 365, no. 1851, pp. 493–514, Feb. 2006.
[53] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, no. 34, USA, 1996, pp. 226–231. [Online]. Available: https://www.osti.gov/biblio/421283
[54] M. Ahmed and A. Naser, "A novel approach for outlier detection and clustering improvement," in *Proc. IEEE 8th Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2013, pp. 577–582.
[55] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti, "Improving k-means by outlier removal," in *Proc. Scandin. Conf. Image Anal.* Berlin, Germany: Springer, 2005, pp. 978–987, doi: 10.1007/11499145_99.
[56] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, p. 3, 2012.
[57] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2009.
[58] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification cost-sensitive boosting," in *Proc. Int. Conf. Mach. Learn.*, vol. 99, 1999, pp. 97–105.
[59] T.-Y. Liu, "Learning to rank for information retrieval," *Found. Trends Inf. Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[60] A. M. Bianco, M. G. Ben, E. J. Martinez, and V. J. Yohai, "Outlier detection in regression models with ARIMA errors using robust estimates," *J. Forecasting*, vol. 20, no. 8, pp. 565–579, Dec. 2001.
[61] M. C. Chuah and F. Fu, "ECG anomaly detection via time series analysis," in *Proc. Int. Symp. Parallel Distrib. Process. Appl.* Berlin, Germany: Springer, 2007, pp. 123–135, doi: 10.1007/978-3-540-74767-3_14.
[62] S. Zhang, A. Chakrabarti, J. Ford, and F. Makedon, "Attack detection in time series for recommender systems," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2006, pp. 809–814.
[63] L. D. Iverson, "Inductive system health monitoring," in *Proc. IC-AI*, 2004, pp. 605–611.
[64] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multi-dimensional time-series with support for multiple distance measures," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2003, pp. 216–225.
[65] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences," *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 259–289, 1997.
[66] B. Rossi, S. Chren, B. Buhnova, and T. Pitner, "Anomaly detection in smart grid data: An experience report," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2016, pp. 2313–2318.
[67] K. Golmohammadi and O. R. Zaiane, "Time series contextual anomaly detection for detecting market manipulation in stock market," in *Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2015, pp. 1–10.
[68] N. Laptev, S. Amizadeh, and I. Flint, "Generic and scalable framework for automated time-series anomaly detection," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 1939–1947.
[69] L. Zhu and N. Laptev, "Deep and confident prediction for time series at Uber," in *Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW)*, Nov. 2017, pp. 103–110.
[70] M. Munir, S. A. Siddiqui, A. Dengel, and S. Ahmed, "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series," *IEEE Access*, vol. 7, pp. 1991–2005, 2018.
[71] C. Zhang, S. Li, H. Zhang, and Y. Chen, "VELC: A new variational AutoEncoder based model for time series anomaly detection," 2019, *arXiv:1907.01702*.
[72] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Y. Chen, B. Hu, N. Begum. (2018). *The UCR Time Series Classification Archive. 2018*. [Online]. Available: https://www.cs.ucr.edu/~eamonn/time_series_data_
[73] W. W. Wei, "Time series analysis," in *The Oxford Handbook of Quantitative Methods in Psychology*, vol. 2. Reading, MA, USA: Addison-Wesley, 2006.
[74] R. H. Jones, "Exponential smoothing for multivariate time series," *J. Roy. Stat. Soc. B, Methodol.*, vol. 28, no. 1, pp. 241–251, 1966.
[75] S. Haykin, *Kalman Filtering and Neural Networks*, vol. 47. Hoboken, NJ, USA: Wiley, 2004.
[76] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing*. Springer, 2009, pp. 1–4.
[77] R. S. Tsay, "Time series and forecasting: Brief history and future research," *J. Amer. Stat. Assoc.*, vol. 95, no. 450, pp. 638–643, Jun. 2000.
[78] N. Meinshausen, "Quantile regression forests," *J. Mach. Learn. Res.*, vol. 7, pp. 983–999, Jun. 2006.
[79] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[80] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002.
[81] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.
[82] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery*. Berlin, Germany: Springer, 2002, pp. 15–27, doi: 10.1007/3-540-45681-3_2.
[83] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. ACM sigmod Rec.*, no. 2, 2000, pp. 93–104.
[84] B. Scholköpf and A. Smola, "Support vector machines, regularization, optimization, and beyond," in *Learning With Kernels*. MIT Press, 2002. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?bknumber=6267332
[85] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019.
[86] S. Das, M. R. Islam, N. K. Jayakodi, and J. R. Doppa, "Active anomaly detection via ensembles: Insights, algorithms, and interpretability," 2019, *arXiv:1901.08930*.
[87] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Comput. Surveys*, vol. 54, no. 1, pp. 1–41, Apr. 2021.
[88] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.
[89] J. Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington, DC, USA: Carnegie Endowment for International Peace, 2020.
[90] C. Cross and R. Gillett, "Exploiting trust for financial gain: An overview of business email compromise (BEC) fraud," *J. Financial Crime*, vol. 27, no. 3, pp. 871–884, Apr. 2020.

**SHWETA TIWARI** is currently pursuing the Ph.D. degree with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. Her research interests include machine learning, deep learning, and information retrieval.

**HERI RAMAMPIARO** is currently the Head of the Department and a Professor at the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. He has previously been the Head of the Data and Artificial Intelligence (DART) Research Group. He has been central in the establishment of the Telenor–NTNU AI Lab, an AI Research Center, NTNU (now Norwegian Open AI Lab), for which he was NTNU's scientific coordinator. His current main research interests include machine learning, information retrieval, and data/text mining.

**HELGE LANGSETH** is currently a Professor of machine learning with the Department of Computer Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. His research interests include probabilistic AI, decision support systems, deep learning, and general machine learning.