Evaluating the Benefits of Using Proactive Transformed-domain-based Techniques in Fraud Detection Tasks

Preprint submitted to Future Generation Computer Systems
(Published version: https://doi.org/10.1016/j.future.2018.10.016)

Roberto Saia and Salvatore Carta[1]

Dipartimento di Matematica e Informatica, Università di Cagliari
Via Ospedale 72, 09124 Cagliari, Italy

Abstract

The exponential growth in the number of E-commerce transactions indicates a radical change in the way people buy and sell goods and services, a new opportunity offered by a huge global market, where they may choose sellers or buyers on the basis of multiple criteria (e.g., economic, logistical, ethical, sustainability, etc.), without being forced to use the traditional brick-and-mortar criterion. If, on the one hand, such a scenario offers an enormous control to people, both at private and corporate level, allowing them to filter their needs by adopting a large range of criteria, on the other hand, it has contributed to the growth of fraud cases related to the involved electronic instruments of payment, such as credit cards. The Big Data Information Security for Sustainability is a research branch aimed to face these issues in relation to the potential implications in the field of sustainability, proposing effective solutions to design safe environments in which the people can operate and by exploiting the benefits related to new technologies. The fraud detection systems are a significant example of such solutions, although the techniques adopted by them are typically based on retroactive strategies, which are incapable of preventing fraudulent events. In this perspective, this paper aims to investigate the benefits related to the adoption of proactive fraud detection strategies, instead of the canonical retroactive ones, theorizing those solutions that can lead toward practical effective implementations. We evaluate two previously experimented novel proactive strategies, one based on the Fourier transform, and one based on the Wavelet transform, which are used in order to move the data (i.e., financial transactions) into a new domain, where they are analyzed and an evaluation model is defined. Such strategies allow a fraud detection system to operate by using a proactive approach, since they do not exploit previous fraudulent transactions, overcoming some important problems that reduce the effectiveness of the canonical retroactive state-of-the-art solutions. Potential benefits and limitations of the proposed proactive approach have been evaluated in a real-world credit card fraud detection scenario, by comparing its performance to

Email address: {roberto.saia,salvatore}@unica.it (Roberto Saia and Salvatore Carta)

that of one of the most used and performing retroactive state-of-the-art approaches (i.e. Random Forests).

Keywords: Business Intelligence, Sustainability, Anomaly Detection

1. Introduction

Nowadays, the *Big Data Analytics for Sustainability* (*BDAS*) represents a crucial research field, since it offers us the opportunity to exploit the new technologies in a smarter way [1, 2], developing more sustainable products and processes.

In the context of the advantages in terms of sustainability offered by the E-commerce environment [3], where the great offer of goods and services allows people to choose those that respect this criterion, helped by the vast amount of information on the Internet, the fraud detection systems represent an instrument around which the interests of many economic entities revolve.

So as it happens in other fields related to technology, even in those where the sustainability represents an essential element, the potential advantages are jeopardized by those who try to take advantage of the new technologies fraudulently. For the aforementioned reasons, one of the most important *BDAS* objectives is the *Big Data Information Security for Sustainability* (*BDISS*). Some *BDISS* areas of great interest are, for instance, those directly related to the security of the adopted platforms (e.g., *Intrusion Detection*, *Fraud Detection*, etc.) and those indirectly related to them (e.g., *Privacy Preserving*, *Cyber Espionage*, etc.).

This paper is focused to one of these important areas, since it faces the problems related to the fraudulent use of the electronic payment instruments, nowadays an essential element for the exchange of goods and services.

Authoritative reports[1] underlined an exponential growth in the fraud losses related to the credit and debit cards, as shown in Figure 1. Several studies[2] have also indicated how the purchases made without authorization and the counterfeits of credit cards represent the *10-15%* of total fraud cases, but the *75-80%* of financial value. Only in the United States, such a problem leads toward an estimated average loss per fraud case of *2* million of dollars.

The aforementioned scenario has generated an increase in research and development investments by private and public entities, with the objective to design more and more effective methods able to face this problem.

It should be observed how the design of effective solutions represents a hard challenge due to several well-known issues, which reduce the capability of the state-of-the-art techniques used in this specific field. The most important issue consists in the fact that the *fraudulent* transactions are typically less than the *legitimate* ones, and such a highly unbalanced data distribution reduces the effectiveness of the machine learning strategies [4]. In addition to this issue there is the scarcity of information that charac-

[1] Nilson Report: https://www.nilsonreport.com/

[2] American Association of Fraud Examiners: http://www.acfe.com

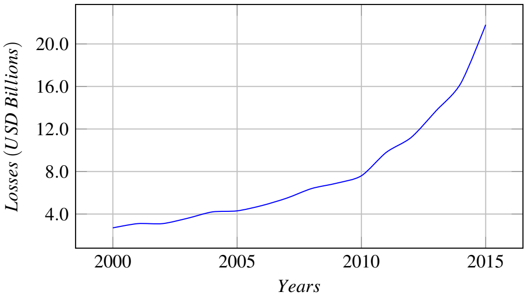

Figure 1: *Annual Global Fraud Losses*

terizes the involved financial transactions, a problem that leads toward an overlapping of the classes of expense of a user.

Nowadays, a fraud detection system can exploit many state-of-the-art techniques in order to evaluate a financial transaction. For instance, it can exploit: *Data Mining* techniques to generate rules from fraud patterns [5]; *Artificial Intelligence* techniques to identify data irregularities [6]; *Neural Networks* techniques to define predictive models [7]; *Signature-based* techniques aimed to maintain a statistical representation of normal account usage for rapid recalculation in real-time [8]; *Fuzzy Logic* techniques to perform a fuzzy analysis for a fraud detection task [9]; *Decision Tree* techniques to reduce the number of misclassifications [10]; *Machine Learning* techniques to define ensemble methods that combine predictions from multiple models [11, 12]; *Genetic Programming* techniques to model and detect fraud through an *Evolutionary Computation* approach [13]; *Statistical Inference* techniques that adopt a flexible Bayesian model for fraud detection [14].

However, it should be observed that regardless of the adopted technique, the principle that is commonly exploited is the detection of outliers in the transactions under analysis, a trivial approach that usually leads toward misclassifications, with all the financial consequences that derive from it (mainly money loss). The reason behind these wrong classifications is the absence of extensive evaluation criteria, since many state-of-the-art techniques are not able to manage some transaction features during the evaluation process (e.g., the non-numeric ones). For instance, *Random Forests* [15], one of the most performing approaches, is not able to manage types of data that involve a large number of categories. For the aforementioned reasons, the evaluation process performed by a fraud detection system should be able to take into account all the information about the transactions under analysis.

Observing how the existence of a relationship of trust between customers and sellers represents one of the most effective proactive strategies able to reduce the fraud issues [16], at the same time we observe how the implementation of automatic proactive strategies is instead a very hard task.

Based on the previous consideration, this paper is aimed to evaluate the benefits related to the adoption of proactive approaches [17], where the analysis of the transaction data is performed in a transformed-domain, instead of a canonical one, by adopting two previously experimented strategies [18, 19], one based on the *Fourier transform* and one based on the *Wavelet transform*.

In the first strategy, the evaluation model is defined in terms of the spectral pattern of a transaction, by processing the information through the *Fourier transformation* [20]. In the second strategy, the evaluation model is defined by following a similar criterion but by processing the information through the *Wavelet transformation* [21].

In both strategies we consider the sequence of values of each transaction feature as a *time series*, moving its representation in the *frequency-domain* by using the *Discrete Fourier Transform* (*DFT*) or in the *time-frequency-domain* by using the *Discrete Wavelet Transform* (*DWT*) process.

The idea is to experiment with different proactive fraud detection strategies in order to investigate about their benefits respect to the canonical retroactive ones, indicating in this way what are the most effective strategies to adopt in the practical implementations.

Many evaluation models adopted in the *Business Intelligence* field are defined on the basis of the *time series*. This happens when it is important to characterize the involved elements on the basis of the time factor [22]. The information extracted from the time series can be exploited in order to perform different tasks, such as those related to the *risk analysis* (e.g., *Credit Scoring* [23] and *Stock Forecasting* [24]) and *Information Security* (e.g., *Fraud Detection* [25] and *Intrusion Detection* [26]) ones.

In the context of the proposed proactive approach, as *time series* we mean something slightly different from the canonical meaning given to it in literature. We refer to it in terms of data used as input in the *DFT* or *DWT* process, thus just in terms of sequence of values that compose a transaction (i.e., *date*, *amount*, and so on), considering them a sequence of discrete-time data taken at successive equally spaced points in time.

In other words, the relationship between *time series* and our fraud detection approach must be sought in the analysis, performed in the frequency domain, of patterns given by the feature values of a transaction. This because our goal is to define an evaluation model able to characterize the legitimate transactions, regardless of the time in which they occur, also because the *time frame* taken into account in our approaches is limited to the *feature space*.

It should be added that the involved transactions are only those related to the previous *legitimate* cases, because we do not consider the previous *fraudulent* ones.

The analysis of the information in the new domain presents some advantages. The first one is the capability for a fraud detection system to include only the previous *legitimate* transactions in the model definition process, which is a proactive approach able to face the *cold-start* issue (i.e., scarcity or absence of *fraudulent* cases during the model definition). Another advantage is related to the consideration that the new data representation reduces the data heterogeneity problem, since the new domain is less influenced by the data variations.

For exemplification reasons, from now on, we will use the term *transformed-domain* to refer to both the data domain obtained by the *Fourier transform* (*frequency-domain*) and the data domain obtained by the *Wavelet transform* (*time-frequency-domain*).

This paper is based on a previous work presented in [25], which scientific contributions were as follows:

(i) definition of the *time series* to use in the Fourier process, on the basis of the past legitimate transactions;

(ii) formalization of the comparison process between the *time series* of an unevaluated transaction and those of the past legitimate transactions, in terms of difference between their frequency magnitude;

(iii) formulation of an algorithm, based on the previous comparison process, able to classify a new transaction as *reliable* or *unreliable*.

Here, our previous work has been completely rewritten and extended, producing the following scientific contributions:

(i) formalization of the procedure used to define the *time series* to use as input in the data transformation process performed by the DFT or DWT strategies, also introducing an alternative method suitable for certain fraud detection contexts;

(ii) formalization of the comparison process between the output obtained at the end of the DFT or DWT process, presenting two different modalities, one based on the *cosine similarity* measured between the entire output vectors of the DFT or DWT processes, and one based on the *punctual comparison* between each element of these output vectors;

(iii) formalization of a generic algorithm able to classify a new transaction as *legitimate* or *fraudulent*, by exploiting the previous comparison process and the DFT or DWT process, also defining its asymptotic time complexity;

(iv) evaluation of our proactive DFT and DWT strategies in terms of general performance and predictive power of their classification model, performed by using two real-world datasets;

(v) evaluation of the advantage and disadvantages related to the adoption of a proactive approach in the context of the fraud detection processes, on the basis of the experimental results.

The paper is organized as follows: Section 2 introduces the background and related work of the scenario taken into account; Section 3 describe the idea behind the proposed proactive approach, introducing the state-of-the-art approach used to evaluate its performance; Section 4 provides a formal notation and defines the problem taken into account; Section 5 describes our proactive approach; Section 6 provides details on the experimental environment, on the used datasets and metrics, as well as on the adopted strategy and selected competitor, presenting the experimental results; some concluding remarks and future work are given in Section 7.

2. Background and Related Work

In this section we first introduce the background of the scenario taken into account by introducing the aspects related to the *big data information security* and the *sustainability*, and by describing the main research problems to be handled. In conclusion, the work related to this field will be described.

2.1. Background

The main challenge of a *Big Data Information Security* (*BDIS*) process is the analysis of huge and heterogeneous data with the goal to protect them against a series of risks such as, for instance, their alteration (*integrity*) or their unauthorized use (*confidentiality*) [27, 28].

It should be observed how in this age of information the risks related to the gathering and use of data are in most of the cases tolerated in view of the great advantages that such operations offer in many fields (e.g., medical, financial, environmental, social, and so on). This kind of paradox has been discussed in literature through several studies, such as those performed in [29].

The main disadvantage of almost all the techniques used to define approaches able to face this type of risk (e.g., alteration or fraudulent use of data) is that they need a considerable number of examples of all possible cases to build their evaluation models (e.g., in the context of a credit card fraud detection system, they need both legitimate and fraudulent examples), precluding the adoption of proactive strategies.

As previously introduced in Section 1, the E-commerce platform allows people to have access to a huge number of goods and services, enabling them to make their own choices on the basis of different criteria. Nowadays, this has been made possible by the coexistence of two factors: a huge E-commerce platform and an equally huge source of information (i.e., Internet).

Through the Internet, people are able to choose sellers and buyers not only on the basis of convenience metrics, but also by following innovative metrics such as, for example, ethical ones.

In this context, dominated by electronic payment instruments, fraud detection systems [30, 25] play a crucial role, since they are aimed to detect the fraudulent financial transactions, allowing people to only get the benefits offered by the E-commerce infrastructure.

The most common problems related to the fraud detection tasks are reported and described in the following.

- **Data Scarcity**: The scarcity of public real-world datasets [31] is the first problem that researchers working in this area have to deal with. It is related to the restrictive policies that regulate the disclosure of information in this area, which they do not allow the operators to provide information about their business activities. Such restrictive rules are related to privacy, competition, or legal reasons. It should be added that not even a release in anonymous form of the data is usually considered acceptable by many financial operators, because even in this form the data may reveal crucial information, such as some vulnerabilities in the E-commerce infrastructure.

- **Non-adaptability**: In the context of a fraud detection system, this is a problem related to the difficulty for the evaluation model to correctly classify new transactions, when they are characterized by patterns differing from those used to train the model. This kind of problem affect both the *supervised* and *unsupervised* fraud detection approaches [32], leading toward misclassifications.

- **Data Heterogeneity**: In the machine learning field, the pattern recognition is a process aimed to assign a label to a given input value. Some common applications of such a process are the classification tasks, where this process is performed in order to classify each value in input into a specific class (within a finite set of classes). It can be exploited in a large number of contexts, thanks to its capability to solve a large number of real-world problems, although its effectiveness is usually affected by the heterogeneity of the involved data. This is a problem described in literature as *naming problem* or *instance identification problem* and it is related to the incompatibility between similar features resulting in the same data being represented differently in different datasets [33, 34]. Given the high level of heterogeneity that characterizes the fraud scenarios (e.g., that related to the credit card transactions), an effective fraud detection system must be able to address the *data heterogeneity* issue.

- **Data Unbalance**: A fraud detection task can be considered as an unbalanced data classification problem, because the examples used to train the evaluation model are typically composed by a large number of *legitimate* cases and a small number of *fraudulent* ones, a data configuration that reduces the effectiveness of the classification approaches [4, 35, 36]. This problem is probably worsened by a *data alteration* in the datasets publicly released by some financial operators, where in order to maintain customer trust in their services, the fraud cases have been intentional reduced, classifying part of them as legitimate. Considering that the canonical approaches of fraud detection operate retroactively, thus they need to train their model by using both classes of examples (i.e., *legitimate* and *fraudulent*), such a problem is commonly faced by preprocessing the dataset in order to obtain an artificial balance of data [37]. This kind of operation can be performed through an *over-sampling* or *under-sampling* method, where in the first case the balance is made by duplicating some of the transactions that are less in number (typically, the *fraudulent* ones), while in the second case it is made by removing some of the transactions that are in greater number (typically, the *legitimate* ones). Some studies demonstrate that the adoption of such methods improves the performance given by the original imbalanced data, also underlining how the *over-sampling* techniques perform better than the *under-sampling* ones [38, 39, 40].

- **Cold-start**: In order to be able to operate properly, machine learning approaches need a significant amount of data to define their evaluation models. While in some contexts this is not a significant issue, in other ones such as, for example, those related to the fraud detection, it represents a big issue. It happens because the examples are characterized by a large number of *legitimate* cases and a small number of *fraudulent* ones, as described in Section 2.1. This configures the so-called *cold-start problem*, i.e., the set of data used to train an evaluation model does not contain enough information about the domain taken into account, making the definition of a reliable evaluation model difficult. In the context taken into account in this paper, this problem arises when the training data is not representative of all the involved classes (*legitimate* and *fraudulent*)) of information [41].

We can summarize all the aforementioned research problems in the need for a fraud detection system to operate by exploiting an evaluation model defined also on the basis of a single class of transactions (i.e., usually the *legitimate* one). Unfortunately, this is not a simple task due several reasons such as the *data unbalance* problem that reduces the performance of the state-of-the-art machine learning techniques commonly used in this field or, even worse, it does not allow us to use them in any way. This is given by the need to train their evaluation model by using examples taken from both the transaction classes (i.e., *legitimate* and *fraudulent*).

Such a problem is further worsened by the lack of real data to use in the process of design and development of new fraud detection approaches (*data scarcity* problem), but also by the difficulty of defining an evaluation model able to work in different contexts (*non-adaptability* problem). Finally, it should be noted how the need to define an evaluation model by using both classes of transactions leads towards the well-known *cold-start* problem.

As detailed in Section 3, the idea around which this paper is born is to define an approach able to train its evaluation model by exploiting a single class of transactions, overcoming in this way the most important problem related with the state-of-the-art approaches that operate in this field (i.e., the *data unbalance*). This solution implicates that such a novel approach must be able to better characterize the class of transaction taken into account during the model definition process, with the side effect to reduce/overcome the other type of problems (i.e., *cold-start*, *data scarcity*, *data heterogeneity*, and *non-adaptability*).

2.2. Related Work

The main goal of a fraud detection system [42, 43] is the evaluation of the new transactions in order to classify them as *legitimate* or *fraudulent*), on the basis of an evaluation model previously defined by exploiting the information gathered by the system during the previous transactions. Premising that the most effective state-of-the-art techniques of fraud detection operate by adopting a retroactive approach, thus they need to train their evaluation models with both the classes of transactions (i.e., *legitimate* and *fraudulent* previous cases), in this section we want to offer an overview of today' s scenario.

The fraud detection approaches can operate by adopting *supervised* or *unsupervised* strategies [44]. By using a *supervised* strategy they exploit the previous *fraudulent* and *non-fraudulent* transactions collected by the system, and they use them to define an evaluation model able to classify a new transaction as *legitimate* or *fraudulent*. In order to perform this task they need to have a sufficient number of examples of both classes, and their recognition capability depends on the known patterns.

By using an unsupervised strategy they instead work by finding anomalies in a transaction under evaluation, in terms of substantial differences in the feature values (wrt the typical values assumed in the past). Considering that a *fraudulent* transaction can be characterized by features with values within their typical range, adopting *unsupervised* strategies in a fraud detection system represents a hard challenge.

The most common approaches [45] used in this context are the *static approach*, the *updating approach*, and the *forgetting approach*, whose features are as follows:

- the *static approach* represents the most common way to operate in order to detect *fraudulent* events in a financial data stream related to a credit card activity. By following it, the data streaming is divided into equal size blocks and the evaluation model is trained by using a limited number of initial and contiguous blocks;

- the *updating approach* adopts instead a different modality, since at each new block the evaluation model is updated by training it with a certain number of latest and contiguous blocks;
- the *forgetting approach* represents another modality, where the evaluation model is updated when a new block appears, and this operation is performed by using all the previous *fraudulent* transactions, along with the *legitimate* transactions present in the last two blocks only.

The models obtained through these approaches can be directly used for the evaluation process or they can be combined in order to define a biggest model of evaluation.

It should be noted that all the aforementioned approaches present some limitations, as reported in the following:

- the *static approach* is not well capable of modeling the users behavior, since its model is defined by using a limited number of blocks;
- the *updating approach* is not able to operate with small classes of data, because its model can be defined only by using a certain number of new blocks;
- the *forgetting approach* presents a high computational complexity, since its model is updated at each new block.

In addition, there are some common issues to overcome that reduce the effectiveness of all these approaches, as described in Section 2.1.

3. Proposed Approach and Competitor

The objective of our approach can be reached by adopting two different strategies, one based on the *Fourier Transform* and one based on the *Wavelet Transform*. This section describes both the aforementioned strategies, discussing at the end their implementation in a fraud detection system.

3.1. Input Data: Time Series Definition

A *time series* usually refers to a series of values acquired by measuring the variation in time of a specific data type (i.e., temperature, amplitude, and so on).

In our approaches we consider as *time series* the sequence of values assumed by the transaction features in the datasets taken into account, introducing also a different modality where the *time series* are defined in terms of sequence of values assumed by each single transaction feature in the dataset domain. This last modality is suitable when we need to model the behavior of a single transaction features, instead of that of all features in the context of a transaction (i.e., how it happens in the considered credit card fraud detection task).

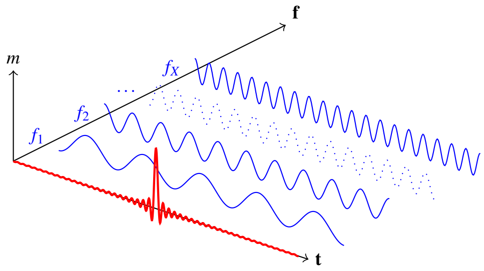

Figure 2: *Time and Frequency Domains*

3.2. Strategy 1: Fourier Transform

The idea behind this first strategy is to perform the evaluation process in a *frequency-domain*, by defining the evaluation model in terms of frequency components. Such a operation is performed by considering the sequence of values assumed by the transaction features as a *time series*, moving its analysis from the canonical domain to a new *transformed-domain* .

The result is a spectral pattern composed by the frequency components, as shown in Figure 2, where t, f, and m, respectively stand for *time*, *frequency*, and *magnitude*. Without claiming to formally show the theoretical concepts related to time and frequency domain, such a figure is aimed to exemplify them by showing how a signal in the time domain can be decomposed in its frequency components.

We made this by recurring to the *Discrete Fourier Transform* (*DFT*) [20], whose formalization is shown in Equation 1, where: X denotes the frequency domain representation of the *time series* x_n (with $n = 0, 1, \dots, N-1$), the formula produces one complex number X_k for each k, k denotes the *k-th* frequency component (with $k = 0, 1, \dots, N-1$), N denotes the number of samples in the *time series* x, n is the *n-th* sample in the time domain, and j denotes the imaginary unit.

$$X_k \overset{\text{def}}{=} \sum_{n=0}^{N-1} x_n \cdot e^{-j2\pi kn/N} \tag{1}$$

The result is a set of sinusoidal functions, each of them related to a specific frequency component. We can return to the original time domain by using the *inverse Fourier transform* reported in Equation 2.

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{j2\pi kn/N} \tag{2}$$

A periodic wave is characterized by a frequency f and a wavelength λ (i.e., the distance in the medium between the beginning and end of a cycle $\lambda = \frac{w}{f_0}$, where w stands for the wave velocity), which are defined by the repeating pattern. Their fundamental period τ is the period where the wave values were taken and sr denotes their number over this time (i.e., the acquisition frequency).

Such a period starts with the value assumed by the feature in the oldest transaction of the set T_+ and it ends with the value assumed by the feature in the newest transaction, thus we have that $sr = |T_+|$; the sample interval si is instead given by the fundamental period τ divided by the number of acquisition, i.e., $si = \frac{\tau}{|T_+|}$.

Assuming that the time interval between the acquisitions is constant, on the basis of the previous definitions applied in the context of this paper, the considered non-periodic wave is given by the sequence of values assumed by each distinct feature $v \in V$ that characterize the transactions in the set T_+ (i.e., the past *legitimate* transactions), and this sequence of values represents the *time series* taken into account in the *DFT* process.

The *transformed-domain* representation, obtained by the *DFT*, process gives us information about the magnitude and phase of the signal at each frequency. Denoting as x the output of the process, it represents a series of complex numbers, where x_r is the real part and x_i is the imaginary one (i.e., we have that $x = (x_r + ix_i)$).

Premising that the magnitude can be calculated by using $|x| = \sqrt{(x_r^2 + x_i^2)}$ and that the phase can be calculated by using $\varphi(x) = \arctan\left(\frac{x_i}{x_r}\right)$, in the context of the presented strategy we will only take into account the frequency magnitude.

We use the *Fast Fourier Transform* (*FFT*) algorithm in order to perform the Fourier transformations, since it allows us to rapidly compute the *DFT* by factorizing the input matrix into a product of sparse (mostly zero) factors. This is a largely used algorithm because it is able to reduce the computational complexity of the process from $O(n^2)$ to $O(n \log n)$ (where n denotes the data size).

3.3. Strategy 2: Wavelet Transform

The idea behind this second strategy is to move the evaluation process from the canonical domain to a new *time-frequency-domain* by exploiting the *Discrete Wavelet Transformation* (*DWT*) [46, 47]. In more detail, we use the *DWT* process in a *time series* data mining context.

The evaluation of the transactions in the new domain offered by the *DWT* leads toward interesting advantages. Such a process transforms a *time series* by exploiting a set of functions named *wavelets* [47], and in literature it is usually performed in order to reduce the data size (e.g., in the image compression tasks) or to reduce the data noise (e.g., in the filtering tasks). The wavelets are mathematical functions that allow us to decompose the original data into different frequencies, then they move the data representation from the time domain (sequence of transaction feature values) to a new domain where the data is represented both in terms of time and scale.

The so-called *time-scale multiresolution* offered by the *DWT* represents an important aspect of this process, since it allows us to observe the original *time series* from different points of view, each containing interesting information on the original data. As frequency we mean the number of occurrences of a value in a *time series* over a unit of time and as scale we mean the time interval that characterize the *time series*. The capability in the new domain to observe the data by using multiple scales (multiple resolution levels), allows us to define a more stable and representative model of the transactions, wrt the canonical approaches at the state of the art.

The process of transformation operated by the *DWT* is different from that carried out by the aforementioned *Fourier Transform* process, since it is characterized by a constant resolution for all the frequencies, differently from DWT that analyzes the data at multiple resolution for different frequencies. Formally, a *Continuous Wavelet Transform* (CWT) is defined as shown in Equation 3, where $\psi(t)$ represents a continuous function in both the time and frequency domain (called *mother wavelet*) and the $*$ denoting the complex conjugate.

$$X_w(a,b) = \frac{1}{|a|^{1/2}} \int_{-\infty}^{\infty} x(t)\psi^* \left(\frac{t-b}{a}\right) dt \tag{3}$$

Given the impossibility to analyze the data by using all wavelet coefficients, it is usually sufficient to consider a discrete subset of the upper half-plane to be able to reconstruct the data from the corresponding wavelet coefficients The considered discrete subset of the half-plane are all the points $(a^m, na^m b)$, where $m, n \in Z$, and this allows us to define the so-called *child wavelets* as shown in Equation 4.

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a^m}} \psi \left(\frac{t-nb}{a^m}\right) \tag{4}$$

The use of small scales (i.e., that corresponds to large frequencies, since the scale is given by the formula $\frac{1}{frequency}$) compresses the data, giving us an overview of the involved information, while large scales (i.e., low frequencies) expand the data, offering a detailed analysis of the information. On the basis of the characteristics of the wavelet transformation, although it is possible to use many basis functions as *mother wavelet* (e.g., *Daubechies*, *Meyer*, *Symlets*, *Coiflets*, etc), for the scope of our proactive approach we decided to use one of the simplest and oldest formalization of wavelets, the *Haar wavelet* [48]. We made this choice because the *Haar wavelet* has the capability of measuring the contrast directly from the responses of low and high frequency sub-bands. This mother wavelet is shown in Equation 5.

$$\psi(t) = \begin{cases} 1, & 0 \leq t > \frac{1}{2} \\ -1, & \frac{1}{2} \leq t < 1 \\ 0, & otherwise \end{cases} \tag{5}$$

For exemplification purposes, considering a *time series* $TS = \{ts_1, ts_2, \ldots, ts_N\}$, for instance $TS = \{8,5,6,7,5,4,6,5\}$ (then with $|TS| = N = 8$), the transformation operated by using the pyramid algorithm of *Haar wavelet* in order to obtain a representation of data based on the average, gives the values reported in Equation 6 as result.

$$\psi(TS) = \{6.5, 6.5, 4.5, 5.5\} \tag{6}$$

The result is obtained by following the criterion shown in the following Equation 7.

$$\frac{ts_2 + ts_1}{2}, ts_2 = \frac{ts_4 + ts_3}{2}, ts_3 = \frac{ts_6 + ts_5}{2}, ts_4 = \frac{ts_8 + ts_7}{2} \tag{7}$$

We can apply the *Haar wavelet* function on the *time series* multiple times, reducing the result length according to the sequence $\frac{N}{2}, \frac{N}{4}, \frac{N}{8}$, and so on. Such a process reduces the level of detail and increases the overview on the data. The *Haar wavelet* function assumes that the length of the input is 2^n, with $n > 1$. When this is not possible, other solutions can be used to overcome this problem, e.g., the *Ancient Egyptian Decomposition* process [49].

The previous example is related only to the *low-pass* process (sum of feature values) performed by the *Haar wavelet*, because a complete application of such a wavelet includes also a *high-pass* process (subtraction of feature values). This was done according to the experiments performed in [18, 19], which show that the results are the same by using only the *low-pass* process, and this also reduces the computational load.

3.4. Implementation

The choice to operate in a transformed domain depends on the need to adopt a strategy able to define an evaluation model on the basis of a single class of data, thus being capable of well characterizing this class of transactions. The two proposed strategies (i.e., *Fourier Transform* and *Wavelet Transform*) allow us to perform this operation by operating proactively, since they only use the previous *legitimate* transactions and they are able to well characterize them by moving the analysis in new data domains (respectively,*frequency-domain* and *time-frequency-domain*).

As detailed explained later in Section 5.2, the *Fourier Transform* has been chosen as first strategy of our proactive approach in order to take advantage of two its interesting properties [50]: the *phase invariance* and the *homogeneity*. Indeed, the *phase invariance* property allows a fraud detection system to detect in a transaction a specific data configuration (feature pattern), regardless of the position of the feature values. It represents an important advantage in the fraud detection tasks, since this capability can reduce the misclassifications, by detecting certain behaviors even when they shift along the feature space. The aforementioned mechanism is further improved by the *homogeneity* property, which allows a fraud detection system to distinguish between two transactions characterized by the same feature pattern but with different feature values.

The *Wavelet Transform* [47], explained in detail in Section 5.2, is the second strategy chosen for our proactive approach. Differently from the *frequency-domain* offered by the *Fourier Transform*, its *time-frequency-domain* allows us to both reduce the data dimensionality (*multiresolution approximation* property) and analyze the data by using multiple resolution levels (*multiresolution analysis* property). These two properties are exploited to reduce the computational complexity related to the fraud detection process and to obtain different points of view on data (e.g., an overview or a detailed view).

The information presented in Table 1 summarize the characteristics of both strategies, reporting also information about their the *asymptotic time complexity*, which determination is explained in detail in Section 5.3.2.

Table 1: Approach Strategies

Approach strategy	Transformed domain	Time complexity	Exploited properties	Implementation type
Fourier transform	*Frequency*	$O(n\ log\ n)$	*Phase invariance, Homogeneity*	*Fast Fourier Transform*
Wavelet transform	*Time-frequency*	$O(n)$	*Multiresolution approximation, Multiresolution analysis*	*Haar Wavelet*

3.5. Competitor

Taking into account that the most effective fraud detection approaches at the state of the art need to train their model by using both the *fraudulent* and *legitimate* previous cases, in this paper we do not compare our approach to many of them, limiting the comparison to only one of the most used and effective ones, being *Random Forests* [15]. The *Random Forests* approach represents one of the most common and powerful state-of-the-art techniques for data analysis, because in most of the cases it outperforms the other ones [51, 35, 52].

It consists in an ensemble learning method for classification and regression based on the construction of a number of randomized decision trees during the training phase. The conclusion are inferred by averaging the obtained results and this technique can be used to solve a wide range of prediction problems, with the advantage that it does not need any complex configuration, because it only requires the adjustment of two parameters: the number of trees and the number of attributes used to grow each tree.

Our aim is to prove that through our proactive approach it is possible to define effective evaluation models built by using only a class of transactions (i.e., the *legitimate* one), granting several advantages.

4. Preliminaries

This section provides the formal notation adopted in this paper and some basic assumptions, as well as the formal definition of the faced problem.

4.1. Formal Notation

The formal notation adopted in this paper is reported in Table 2. It should be observed that a transaction can only belong to one class $c \in C$.

Table 2: Formal Notation

Notation	Description	Note
$T = \{t_1, t_2, \dots, t_N\}$	Set of classified transactions	
$T_+ = \{t_1, t_2, \dots, t_K\}$	Subset of legitimate transactions	$T_+ \subseteq T$
$T_- = \{t_1, t_2, \dots, t_J\}$	Subset of fraudulent transactions	$T_- \subseteq T$
$V = \{v_1, v_2, \dots, v_M\}$	Set of transaction features	
$\hat{T} = \{\hat{t}_1, \hat{t}_2, \dots, \hat{t}_U\}$	Set of unclassified transactions	
$C = \{legitimate, fraudulent\}$	Set of possible classifications	
$F = \{f_1, f_2, \dots, f_X\}$	Output of DFT or DWT process	

4.2. Problem Definition

Denoting as Ξ the process of comparison between the DFT (or DWT) output of the *time series* in the set T_+ (i.e., the sequence of feature values in the previous *legitimate* transactions) and the DFT (or DWT) output of the *time series* related to the unevaluated transactions in the set $\hat{T}$ (processed one at a time), the objective of our proactive approach is the classification of each transaction $\hat{t} \in \hat{T}$ as *legitimate* or *fraudulent*.

Defining a function $EVAL(\hat{t},\Xi)$ that performs this operation based on our approach, returning a boolean value β (*0=misclassification*, *1=correct classification*) for each classification, we can formalize our objective function (Equation 8) in terms of maximization of the results sum.

$$\max_{0\leq\beta\leq|\hat{T}|} \beta = \sum_{u=1}^{|\hat{T}|} EVAL(\hat{t}_u,\Xi) \tag{8}$$

5. Proposed Approach

The proposed approach was implemented by performing the steps listed below and detailed in the Sections 5.1, 5.2, and 5.3.

1. **Data Definition**: definition of the *time series* to use as input in the DFT or DWT process, in terms of sequence of values assumed by the transaction features;

2. **Data Processing**: conversion of the *time series* obtained in the previous step into a *transformed-domain* by using the DFT or DWT process;

3. **Data Evaluation**: formalization of the algorithm able to classify a new transaction as *legitimate* or *fraudulent* on the basis of a comparison process made in the *transformed-domain*.

5.1. Data Definition

As previously introduced in Section 3.1, a *time series* is a sequence of data points stored by following the time order and, in most of the cases, it is a sequence of discrete-time data measured at successive equally spaced points in time.

In the context of our approach, we considered as *time series* (*ts*) the sequence of values $v \in V$ assumed by the features of the transactions in T_+ and $\hat{T}$, as shown in Equation 9 and Equation 10.

$$T_+ = \begin{vmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,M} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{K,1} & v_{K,2} & \cdots & v_{K,M} \end{vmatrix} \qquad \hat{T} = \begin{vmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,M} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ v_{U,1} & v_{U,2} & \cdots & v_{U,M} \end{vmatrix} \tag{9}$$

$$\begin{array}{l} ts(T_+) = (v_{1,1},v_{1,2},\ldots,v_{1,M}),(v_{2,1},v_{2,2},\ldots,v_{2,M}),\cdots,(v_{K,1},v_{K,2},\ldots,v_{K,M}) \\ ts(\hat{T}) = (v_{1,1},v_{1,2},\ldots,v_{1,M}),(v_{2,1},v_{2,2},\ldots,v_{2,M}),\cdots,(v_{U,1},v_{U,2},\ldots,v_{U,M}) \end{array} \tag{10}$$

The *time series* related to an item $\hat{t} \in \hat{T}$ will be compared to the *time series* related to all the items $t_+ \in T_+$, by following the criteria explained in the next steps.

An alternative method defines the *time series* in terms of sequence of values assumed by each transaction feature $v \in V$ in the set T_+ and $\hat{T}$, as shown in Equation 11. Such a different modality is suitable when the aim of the evaluation model is to detect atypical values in a single feature, rather than in the whole set of features. Since we need to detect atypical values in the whole set of features, we adopt the first modality, i.e., that shown in Equation 10.

$$\begin{array}{l} ts(T_+) = (v_{1,1}, v_{2,1}, \ldots, v_{K,1}), (v_{1,2}, v_{2,2}, \ldots, v_{K,2}), \cdots, (v_{1,M}, v_{2,M}, \ldots, v_{K,M}) \\ ts(\hat{T}) = (v_{1,1}, v_{2,1}, \ldots, v_{U,1}), (v_{1,2}, v_{2,2}, \ldots, v_{U,2}), \cdots, (v_{1,M}, v_{2,M}, \ldots, v_{U,M}) \end{array} \tag{11}$$

5.2. Data Processing

The *time series* defined in the previous step are here processed in order to move their representation to the *transformed-domain*, by using the *DFT* or *DWT* process.

In a preliminary study we compared some patterns in the time domain (i.e., the *time series*) to their representation in the *transformed-domain*. Without going deep into the merits of the formal characteristics of *Fourier* and *Wavelet* transformations, thus limiting our analysis to the context taken into account, we underlined the properties described below:

5.2.1. Exploited Fourier Properties

1. **Phase invariance:** the first property, shown in Figure 3, demonstrates that there are not variations in the spectral pattern in case of a value translation[3]. More formally, it is one of the *phase properties* of the Fourier transform [50], i.e., a shift of a *time series* in the time domain leaves the magnitude unchanged in the *transformed-domain* [50]. It means that the representation in the *transformed-domain* allows us to detect a specific pattern, regardless of the position of the values assumed by the transaction features that originate it;

2. **Magnitude correlation:** the second property, shown in Figure 4, instead proves the existence of a direct correlation between the values assumed by the features in the time domain and the corresponding magnitudes assumed by the spectral components in the *transformed-domain*. More formally, it is the *homogeneity property* of the Fourier transform [50], i.e., when the magnitude is altered in one domain, it is altered by the same entity in the other domain[4]. This ensures that the proposed approach is able to evaluate the differences in terms of feature values, i.e., it is able to differentiate identical spectral patterns on the basis of the values assumed by their transaction features;

3. **Additivity quality:** another interesting property, shown in Figure 5, allows us to define patterns able to represent particular user behaviors, by simply adding the *time series* related to the involved transactions. More formally, it represents the *additivity property* of the Fourier transform [50], i.e., to the addition in the time domain corresponds an addition in the frequency domain. It means that we can merge two patterns in the time domain, without losing information in the spectral pattern representation.

[3] A translation in time domain corresponds to a change in phase in the frequency domain.

[4] Scaling in one domain corresponds to scaling in the other domain

By adopting the *Fourier* strategy, in this step we move the *time series* of the transactions to the *transformed-domain* by using a DFT process performed through the FFT algorithm introduced in Section 3.2 and Section 3.3.. Basically, we extract the spectral pattern of each transaction by processing the related *time series* defined in the previous step.

5.2.2. Exploited Wavelet Properties

1. **Dimensionality reduction:** the DWT process represents an effective method for the *time series* data reduction, since the orthonormal transformation operated reduces the dimensionality of a *time series*, providing a compact representation of data, which however preserves the original information in its coefficients (multiresolution approximation). By exploiting this property, a fraud detection system can reduce the computational complexity of the involved processes;

2. **Multiresolution analysis:** applied on the *time series* context, the DWT allows us to define separate *time series* on the basis of the original one, distributing the information in these new representations of data in terms of the wavelet coefficient. The most important aspect of such transformations is that the DWT process performs an orthonormal transformation, preserving the original information, allowing us to restore the original data representation. A fraud detection system can exploit this mechanism in order to detect rapid changes in the data under analysis, observing the *data series* under two different points of view (i.e., types of wavelet coefficient), an approximated and a detailed one. The approximate point of view provides an overview on the data, while the detailed point of view provides information useful to evaluate data changes.

By using the *Wavelet* strategy, in this step we transform the original *time series* given by the sequence of values assumed by the transaction features (as explained in Section 5.1) by performing the *Haar wavelet* process described in Section 3.3. We apply the *Haar wavelet* function on the *time series* one time and a wavelet coefficient that leads towards an approximation of the data (i.e., $\frac{N}{2}$) is preferred in order to define a more stable model (i.e., less influenced by the data heterogeneity) for the evaluation of the new transactions. This happens because the differences between the transaction feature values are reduced during the average process performed by the *Haar wavelet* (Equation 7).

5.3. Data Evaluation

The process of evaluation of a new transaction is performed by comparing the DFT or DWT outputs of the previous *legitimate* transactions to those of the transactions to evaluate.

For each transaction $\hat{t} \in \hat{T}$ we compare its *transformed-domain* representation $F(\hat{t})$ (i.e., the series of values $f \in F$) to the *transformed-domain* representation $F(t_+)$ of each *legitimate* previous transaction $t_+ \in T_+$.

The comparison process can be done in the *transformed-domain* (i.e., DFT or DWT outputs vectors) by using one of the two different methods described in the following:

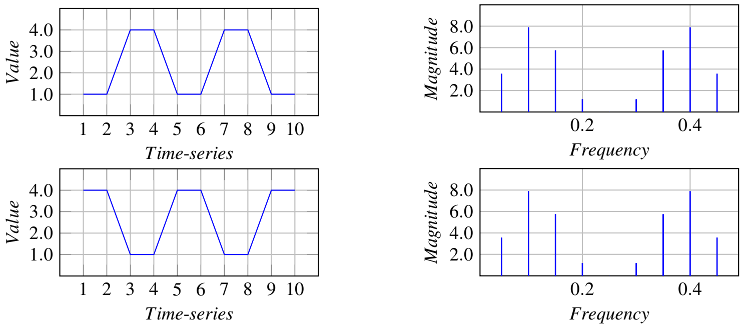

Figure 3: *Fourier* : *Phase Invariance*

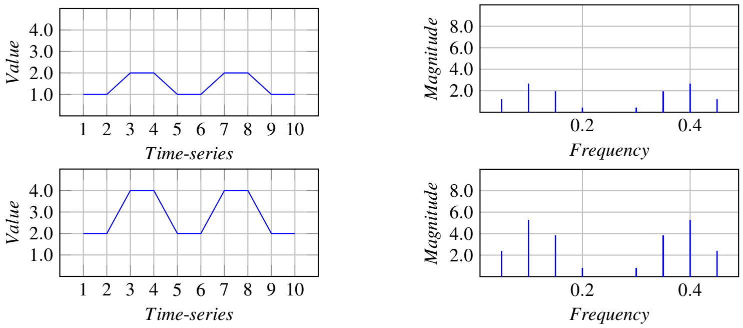

Figure 4: *Fourier* : *Magnitude Correlation*

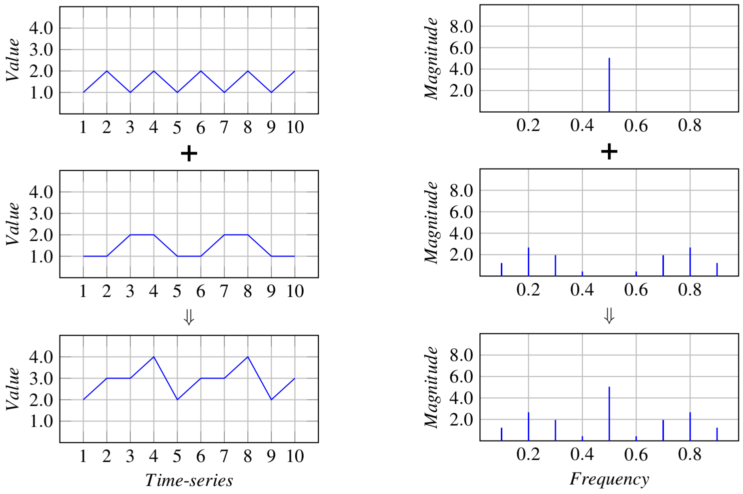

Figure 5: *Fourier* : *Additivity Quality*

1. the first method is based on the *cosine similarity*, a well-known metric described in Section 6.3.1. This first method is suitable when we need to evaluate the similarity between transactions in a global manner, thus by jointly evaluating the behavior of all the elements that compose the output vectors of the DFT or DWT process;

2. the second method is based on the *punctual comparison* between the values assumed by each element of the output vectors, with regard to the minimum or maximum value assumed by the element in the dataset (the result will be a boolean value, then *0* or *1*). The similarity is evaluated with respect to a threshold, e.g., a transaction is considered similar to another one, when the sum of the comparison results of all the elements is above the $\frac{|F|}{2}$ value. Such a method is suitable when we need to evaluate the similarity on the basis of the behavior of each single feature (e.g., in some *Intrusion Detection Systems* [53]).

For the aforementioned considerations, in our strategies we adopt the first method (i.e., *cosine similarity*), as shown in Equation 12, where Δ represents the similarity value in terms of *cosine similarity*, α is a threshold value experimentally defined in Section 6.4, and c is the resulting classification.

We repeat this process by using the transaction to evaluate $\hat{t}$ and all the transactions $t_{+} \in T_{+}$, obtaining the final classification of the transaction by averaging over these comparisons.

$$\Delta = cos(F(t), F(\hat{t})),\ with\ c = \begin{cases} \Delta \geq \alpha, & \text{legitimate} \\ \Delta < \alpha, & \text{fraudulent} \end{cases} \tag{12}$$

5.3.1. Algorithm

The final classification of a new transaction $\hat{t}$, which takes into account all the comparisons (Equation 12) between the transaction $\hat{t}$ and all the transactions in T_{+}, is performed by using the Algorithm 1.

This process takes as input the set T_{+} of past *legitimate* transactions, a transaction $\hat{t}$ to evaluate, and the threshold value α to use in the spectral pattern comparison process (i.e., in the context of the *cosine similarity* evaluation). It returns as output a boolean value that indicates the $\hat{t}$ classification (i.e., *true*=*legitimate* or *false*=*fraudulent*).

From *step 1* to *step 16* we process the unevaluated transaction $\hat{t}$, by starting with the definition of the *time series* related to the transaction $\hat{t}$ (*step 2*), moving it in the *transformed-domain* (*step 3*).

In the *steps* from *4* to *8*, we compare in the *transformed-domain* the transaction $\hat{t}$ to that of each transaction $t_{+} \in T_{+}$ (obtained at the *steps 5* and *6*), adding the result (i.e., the *cosine similarity* value) to the variable *cos* (*step 7*).

The average of the final value of the variable *cos* (*step 9*) is compared to the threshold value α (*steps* from *10* to *14*)), and the final classification of the transaction $\hat{t}$, returned by the algorithm at the *step 15*, depends on the result of this operation.

5.3.2. Complexity

Here we evaluate the *asymptotic time complexity* of the Algorithm 1. This represents an information that allows us to evaluate the performance of the proposed ap-

Algorithm 1 *Transaction evaluation*

```
Input: $T_+$=Legitimate previous transactions, $\hat{t}$=Unevaluated transaction, $\alpha$=Threshold value
Output: $\beta$=Classification of the transaction $\hat{t}$
 1: procedure TRANSACTIONEVALUATION($T_+$, $\hat{t}$)
 2:     $ts1 \leftarrow getTimeseries(\hat{t})$
 3:     $sp1 \leftarrow getTransformedDomain(ts1)$
 4:     for each $t_+$ in $T_+$ do
 5:         $ts2 \leftarrow getTimeseries(t_+)$
 6:         $sp2 \leftarrow getTransformedDomain(ts2)$
 7:         $cos \leftarrow cos + getCosineSimilarity(sp1, sp2)$
 8:     end for
 9:     $avg \leftarrow \frac{cos}{|T_+|}$
10:     if $avg > \alpha$ then
11:         $\beta \leftarrow true$
12:     else
13:         $\beta \leftarrow false$
14:     end if
15:     return $\beta$
16: end procedure
```

proach in the context of a *real-time system* [54], a scenario where the *response-time* represents a primary aspect. According to the *Big O notation*, we determinate it on the basis of the following observations:

(i) the Algorithm 1 performs a single loop (from *step 4* to *step 8*) and other simple operations of *comparisons* and *assignations*;

(ii) the loop recalls three functions (*getTimeseries*, *getTransformedDomain*, and *getCosineSimilarity*) and performs three *assignment operations* (*steps 5*, *6*, and *7*);

(iii) the complexity of the aforementioned operations is, respectively, $O(n)$ (*getTimeseries*), $O(n\ log\ n)$ (*getTransformedDomain*), $O(n^2)$ (*getCosineSimilarity*), and $O(1)$ (*assignment operation*);

(iv) the complexity $O(n\ log\ n)$ assigned to the transformation process (performed at *step 6*) represents the worst case of our two strategies, since the *Discrete Wavelet Transform* process takes only $O(n)$ in certain cases, as compared to $O(nlogn)$ of the *Fast Fourier Transform* process.

The aforementioned observations allow us to determine that the *asymptotic time complexity* of the proposed algorithm is $O(n^2)$, a complexity that can be effectively reduced by parallelizing the process over several machines, e.g., by exploiting large scale distributed computing models such as *MapReduce* [55].

6. Experiments

This section reports information about the experimental environment, the used datasets and metrics, the adopted strategy, as well as the results of the performed experiments. Finally, such experimental results will be analyzed and discussed with the aim to indicate the most effective proactive strategies.

6.1. Environment

The proposed approach was developed in Java, where we use the *JTransforms*[5] library to operate the Fourier transformations, and the *JWave*[6] library to operate the Wavelet transformations.

The preliminary analysis of the ten most performing state-of-the-art approaches, which was aimed to detect the best one to use as competitor during the experiments was performed by using the *Waikato Environment for Knowledge Analysis* (*WEKA*)[7].

The competitor state-of-the-art approach and the metrics used to evaluate its results were instead implemented in *R*[8], by using *randomForest*, *DMwR*, and *ROCR* packages.

The computer used for the experiments was an *Intel Quad-core i7-4510U*, clocked at *2.0 GHz* with *12 GB* RAM and operating system Linux *Debian 8.6* (*Jessie*), kernel version *3.16.0-4-amd64*.

It should be further added that we verified the existence of a statistical difference between the results, by using the independent-samples *two-tailed Student's t-tests* ($p < 0.05$).

6.2. Datasets

The two real-world datasets used in the experiments (i.e., *European Transactions*[9] and *German Credit*[10]) represent two benchmarks in this research field. They are widely adopted by researchers worldwide, taking into account the data scarcity issue previously described in Section 2.1.

We chose them in order to evaluate our proactive approach in two different scenarios in terms of data imbalance and data size. This is possible because the first dataset (i.e., *European Transactions*) is composed by *284,807* transactions with the 0.0017% of frauds, while the second one (i.e., *German Credit*) contains *1,000* transactions with the 30% of frauds,

- **European Transactions (ET)**: This dataset contains the transactions carried out in two days of September 2013, for a total of *492* frauds out of *284,807* transactions. It should be observed how this represents an highly unbalanced dataset [56], considering that the *fraudulent* cases are only the *0.0017%* of all the transactions. For confidentiality reasons, all fields of the dataset have been anonymized, except the *time* (that we do not take into account in the Fourier transformation process) and *amount* features that report, respectively, the number of seconds elapsed between the first transaction in the dataset and the current transaction, and the amount of the credit card transaction. As usual, the last field contains the transaction classification (*0=legitimate* and *1=fraudulent*).

[5] https://sourceforge.net/projects/jtransforms/

[6] https://github.com/cscheiblich/JWave/

[7] http://www.cs.waikato.ac.nz/ml/weka/

[8] https://www.r-project.org/

[9] https://www.kaggle.com/dalpozz/creditcardfraud/

[10] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/

- **German Credit (GC)**: This dataset is composed by *1,000* transactions and *300* of them are frauds. Also in this case it represents an unbalanced dataset, since the *fraudulent* cases are the *30%* of all the transactions. The dataset is released with all the features modified for confidentiality reasons, and we used the version with all numeric features (i.e., without categorical variables). Each transaction is composed by *20* fields, plus a classification field (*1=legitimate* and *2=fraudulent*).

6.3. Metrics

This section introduces the metrics used in the context of this paper.

6.3.1. Cosine Similarity

The *cosine similarity* (*Cosim*) between two non-zero vectors $\vec{v_1}$ and $\vec{v_2}$ is calculated in terms of the cosine angle between them, as shown in the Equation (13).

It represents a widespread measure that allows us to evaluate the similarity between two transaction patterns by comparing the vectors given by the values of their components in the *transformed-domain*.

We chose to adopt this metric since, compared to other similarity metrics, it best captures the differences between two transaction patterns and, in addition, it works better with high dimensional data.

$$Cosim(\vec{v_1},\vec{v_2}) = cos(\vec{v_1},\vec{v_2}) = \frac{\vec{v_1} \cdot \vec{v_2}}{\parallel \vec{v_1} \parallel \cdot \parallel \vec{v_2} \parallel} \tag{13}$$

6.3.2. F-score

The *F-score* is considered an effective performance measure for unbalanced datasets [56]. It represents the weighted average of the *Precision* and *Recall* metrics and it is a largely used metric in the statistical analysis of binary classification, returning a value in a range $[0,1]$, where 0 is the worst value and 1 the best one.

More formally, given two sets $T^{(P)}$ and $T^{(R)}$, where $T^{(P)}$ denotes the set of performed classifications of transactions, and $T^{(R)}$ the set that contains the actual classifications of them, this metric is defined as shown in Equation 14.

$$\begin{array}{l} F\text{-}score(T^{(P)},T^{(R)}) = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recal} \\ \text{with} \\ Precision(T^{(P)},T^{(R)}) = \frac{|T^{(R)} \cap T^{(P)}|}{|T^{(P)}|} \\ Recall(T^{(P)},T^{(R)}) = \frac{|T^{(R)} \cap T^{(P)}|}{|T^{(R)}|} \end{array} \tag{14}$$

6.3.3. AUC

The *Area Under the Receiver Operating Characteristic* curve (*AUC*) is a performance measure used to evaluate the effectiveness of a classification model [57]. Its result is in a range $[0,1]$, where 1 indicates the best performance.

It measures the capability of a binary classifier to discern between two classes of events (in our case, between *legitimate* and *fraudulent* transactions). This is equivalent to the probability that a binary classifier will rank a randomly chosen legitimate transaction higher than a randomly chosen fraudulent one (assuming legitimate ranks higher than fraudulent).

More formally, according to the notation of Section 4.1, given the subset of previous *legitimate* transactions T_+ and the subset of previous *fraudulent* ones T_-, the formalization of the AUC metric is reported in the Equation 15, where Θ indicates all possible comparisons between the transactions of the two subsets T_+ and T_-. It should be noted that the result is obtained by averaging over these comparisons.

$$\Theta(t_+,t_-) = \begin{cases} 1, & if\, t_+ > t_- \\ 0.5, & if\, t_+ = t_- \\ 0, & if\, t_+ < t_- \end{cases} \quad AUC = \frac{1}{|T_+|\cdot|T_-|}\sum_{1}^{|T_+|}\sum_{1}^{|T_-|}\Theta(t_+,t_-) \tag{15}$$

6.4. Strategy

In order to reduce the impact of data dependency, improving the reliability of the obtained results, all the experiments have been performed by using the *k-fold cross-validation* criterion, with *k=10*. Each dataset is divided in *k* subsets, and each *k* subset is used as the test set, while the other *k-1* subsets are used as the training set. The final result is given by the average of all results. All the experiments, including the tuning processes, have been performed by following such a *k-fold cross-validation* criterion.

Before starting the experiments we carried out a study aimed to identify the best value of the threshold parameter α to use in the evaluation process, according to the Equation 12. In order to maintain a proactive strategy, we perform this operation by using only the *legitimate* transactions in the dataset, calculating the average value of the *cosine similarity* related to all pairs of different transactions $t_+ \in T_+$, according to the Algorithm 2. In this process we compare (i.e., in terms of *cosine similarity*) all different pairs of *legitimate* transactions. We perform this operation in the *transformed-domain* of our approach (i.e., obtained by using the *Fourier Transform* or the *Wavelet Transform* strategy), and our aim is to determinate the average value of similarity between two *legitimate* transactions. Such information will be used to detect the potential *fraudulent* transactions.

It takes as input the set T_+ of past *legitimate* transactions and returns the threshold value α to use in the Algorithm 1. The two nested loops that start at *step 2* and at *step 3* select only the different pairs of *legitimate* transactions (*step 4*) in the set T_+. For these pairs (i.e., $t^{'}_+$ and $t^{''}_+$), it calculates their *time series*, which moves in the *transformed-domain* (*steps* from *6* to *9*). Finally, it adds to the *cos* variable the *cosine similarity* calculated between them (*step 10*). The *cosine similarity* average (α) is calculated at *step 14* and this value is returned by the algorithm at the *step 15*.

The evaluation was stopped when the value of α did not present significant variations. In both datasets, the results indicate $\alpha = 0.90$ as the optimal threshold to use in the DFT strategy and $\alpha = 0.91$ as the optimal threshold to use in the DWT strategy.

As mentioned in Section 6.4, the experiments have been performed by following the *k-fold cross-validation* criterion, then the result α represents the average of *k* results obtained by following this criterion.

Such a threshold-based modality has been chosen because the transaction comparison takes place in terms of *cosine similarity*, therefore between two values.

Algorithm 2 *Threshold tuning*

```
Input: $T_+$=Legitimate previous transactions
Output: $\alpha$=Threshold value
procedure GETALPHA($T_+$)
    for each $t'_+$ in $T_+$ do
        for each $t''_+$ in $T_+$ do
            if $t'_+ \neq t''_+$ then
                $evaluations \leftarrow evaluations + 1$
                $ts1 \leftarrow getTimeseries(t'_+)$
                $sp1 \leftarrow getTransformedDomain(ts1)$
                $ts2 \leftarrow getTimeseries(t''_+)$
                $sp2 \leftarrow getTransformedDomain(ts2)$
                $cos \leftarrow cos + getCosineSimilarity(sp1, sp2)$
            end if
        end for
    end for
    $\alpha \leftarrow \frac{cos}{evaluations}$
    return $\alpha$
end procedure
```

It should be noted that the threshold tuning is not to be considered a static process, since it should be repeated periodically on the basis of the operative context in order to get the best performance, preferably during the system downtimes.

6.5. Competitor

As introduced in Section 6.1, we compare our proactive approach to *Random Forest*, because many studies in literature indicates it as the most performing approach for the fraud detection tasks. However, we conducted a preliminary experimentation aimed to test the performance of ten different state-of-the-art approaches, i.e., *Naive Bayes*, *Logit Boost*, *Logistic Regression*, *Stochastic Gradient Descent*, *Multilayer Perceptron*, *Voted Perceptron*, *Random Tree*, *K-nearest*, *Decision Tree*, and *Random Forests*.

We use *AUC* as evaluation metric, since it gives us a measure of the evaluation model effectiveness and the results reported in Table 3 confirm *Random Forests* as the most performing approach.

Table 3: Competitor Approaches Performance

Approach	AUC	Approach	AUC
Naive Bayes	0.889	**Logit Boost**	0.925
Logistic Regression	0.920	**SGD**	0.847
Multilayer Perceptron	0.918	**Voted Perceptron**	0.813
Random Tree	0.851	**K-nearest**	0.864
Decision Tree	0.861	**Random Forests**	0.945

6.5.1. Description

Random Forests has been implemented in *R* language, by using the *randomForest* package. We also used the *DMwR* package to operate the data balancing by following the *Synthetic Minority Over-sampling Technique* (*SMOTE*) [58].

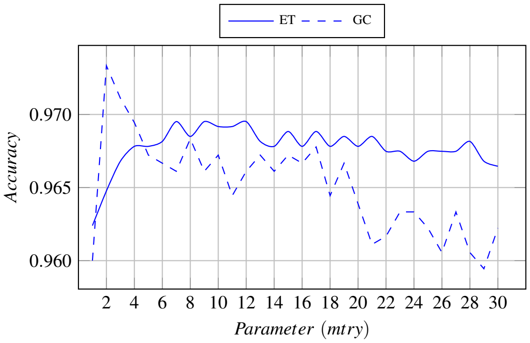

Figure 6: *Random Forests Tuning*

Such a combined approach (i.e., *Random Forests* and *SMOTE*) allows us to evaluate the performance of our proactive approach against one of the best state-of-the-art approach.

To ensure the experiments reproducibility, we used the *R* function *set.seed()* in order to set the seed of the random number generator.

6.5.2. Tuning

In order to maximize the performance of the *RF* approach, we used the *caret R* package in order to detect the optimal value of the *mtry* (number of variables randomly sampled as candidates at each split) parameter[11].

As first step, we preprocess the original dataset by using *SMOTE*, obtaining as result a balanced dataset. In this process *SMOTE* has been configured by setting *perc.over=200*[12] and *perc.under=150*[13]. These two values are those that allow us to balance the *legitimate* and *fraudulent* cases (i.e., by *over-sampling* the minority class).

The results in Figure 6 indicate $mtry = 12$ as optimal value for the *ET* dataset and $mtry = 2$ as optimal value for the *GC* dataset. This because such values are those that lead toward the best value of *Accuracy*, which are, respectively, 0.969% and 0.973%.

It should be noted that all the experiments have been performed by following the *k-fold cross-validation* criterion described in Section 6.4, then the result in Figure 6 represents the average of *k* results.

6.6. Results

The observations and considerations that arise by analyzing the experimental results are reported in this section.

(i) our proactive approach, implemented by following two different strategies, one based on the *Fourier Transform* (*DFT*) and one based on the *Wavelet Transform*

[11] The *RF* performance are strongly related to this parameter
[12] This parameter drives the *over-sampling*
[13] This parameter drives the *under-sampling*

(*DWT*). has been evaluated under two different point of views. The first of them (i.e., *F-score*) allows us to have an overview about the evaluation metrics based on the *confusion matrix*[14], while the second of them (i.e., *AUC*), which is a quality metric based on the analysis of the area under the *ROC*[15] curve, allows us to measure the effectiveness of our evaluation model;

(ii) the first set of experiments was focused on the evaluation of the proposed proactive approach in terms of *F-score*. It is aimed to evaluate its capability to reach a good balance between *Precision* and *Recall* performances, regardless of the size of data and the level of imbalance of them. The results in Figure 7 indicate that both the *DFT* and *DWT* performance are similar to that of *Random Forests* (*RFS*), in the context of the *ET* and *GC* datasets taken into account;

(iii) a good *F-score* performance is obtained despite our approach does not exploit any previous *fraudulent* transaction to train its model, adopting a pure proactive strategy. It means that it can operate without the need to train its model with both classes of transactions (*legitimate* and *fraudulent*);

(iv) the second set of experiments was instead aimed to evaluate our proactive approach in terms of *AUC*. As described in Section 6.3.3, this metric evaluates the predictive power of a classification model and the results in Figure 8 show how our approach is able to achieve performance close to that of *RFS* in the context of both *ET* and *GC* datasets, although it gets better performance with the *ET* one;

(v) such a difference in the *AUC* performance is given by the fact that our approach defines its evaluation model exclusively on the basis of legitimate cases, therefore the greater number of these cases in the *ET* dataset (i.e., *284,315* cases against the *700* legitimate ones of the *GC* dataset) allows it to achieve better performance.

(vi) Both sets of experiments show that the evaluation models (*Fourier-based* and *Wavelet-based*) adopted by our proactive approach reach similar performance in terms of *AUC* in the context of each considered dataset (i.e., respectively, *0.77* and *0.78* in the *ET* dataset and *0.63* and *0.64* in the *GC* dataset), while there are more differences in terms of *Precision* and *Recall* metrics, whose weighted average is shown through the *F-score* metric in Figure 7;

(vii) the differences in terms of *F-score* are related to the fact that the *Wavelet* transformation presents some advantages with regard to the *Fourier* one, since it allows us to analyze the data at multiple resolutions for different frequencies, as described in Section 3.3, offering the possibility to adapt the evaluation model to the operative scenario.

[14] A matrix 2x2 where are reported the number of True Negatives (*TN*), False Negatives (*FN*), True Positives (*TP*), and False Positives (*FP*).

[15] Receiver Operating Characteristic

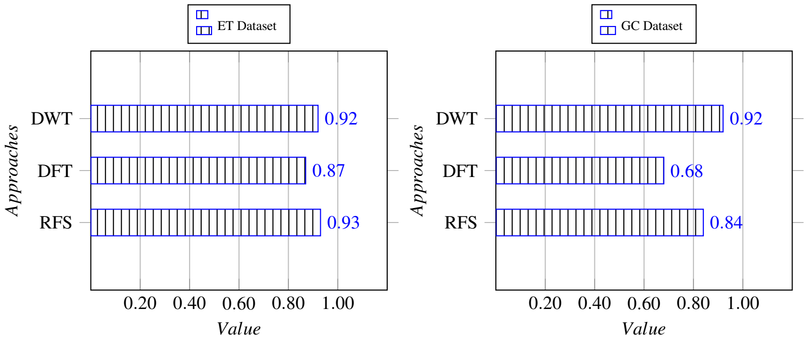

Figure 7: *F-score Performance*

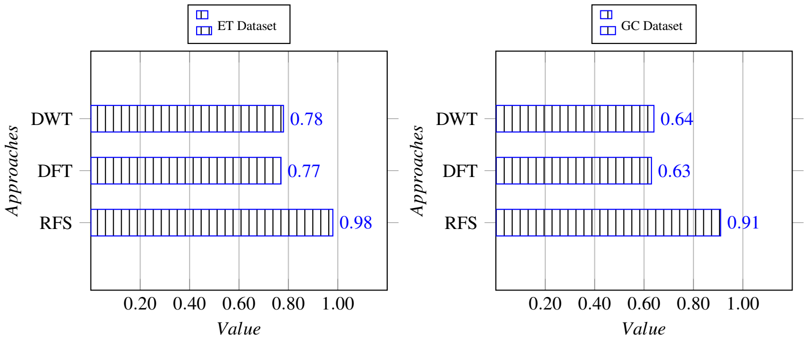

Figure 8: *AUC Performance*

6.7. Discussion

In the light of the aforementioned observations and considerations, we can deduce that the modeling of the transactions made in a *transformed-domain* is able to face the problems related to the *non-adaptability* and the *heterogeneity issues* described in Section 2.1, thanks to the stability offered by the new data representation.

We can also observe how the proactive strategy followed by our approach is able to reduce/overcome the *data imbalance* and *cold-start* issues, which are also described in Section 2.1, since the two proactive strategies used by our approach exploit only a single class of transactions (i.e., the *legitimate* one).

The most important aspect of such a proactivity is represented by the fact that it allows a real-world fraud detection system to operate even in the absence of previous *fraudulent* cases, with all the obvious advantages that derive from it.

With regard to the retroactive and proactive aspect of the fraud detection techniques, it should be evaluated on the basis of the operative scenario. In the scenario taken into account it is in fact reasonable to tolerate that a proactive approach gets worse performance than that of a retroactive one.

This statement is based on the consideration that through a proactive approach a fraud detection system can operate without the need to collect a number of *fraudulent* transactions to use for the model training (used by the retroactive approaches), reducing the economic losses.

It should also be added how the proposed proactive approaches can be considered in the context of hybrid techniques, since their correct classifications can be used in order to improve the effectiveness of the canonical retroactive state-of-the-art approaches, reducing the data unbalance issue. This means that a combined approach, which adopts retroactive and proactive techniques, can be used to design a very effective fraud detection system, where the capabilities of the single approaches are optimized.

The results indicate that the differences between the competitor and our best performing approach (i.e., that based on the *DWT* strategy) are really minimal (Table 4), despite it adopts a pure proactive strategy. This minimum difference in performance must be further reduced in the light of the fact that the misclassifications made by our approach do not necessarily lead toward loss of money, as they are related to both *false positive* and *false negative* cases.

Summarizing, through the adoption of proactive approaches, such as those proposed in this paper, we can contrast the issues discussed in Section 2.1, as their processes do not involve *fraudulent* examples (facing the *data scarcity* and *data unbalance* issues), adopting a *transform-domain-based* model able to well characterize a specific class of transactions (i.e., the *legitimate* one) that results less influenced by the data variation (facing the *non-adaptability* and *data heterogeneity* issues), presenting the positive side effect of solving the *cold-start* issue.

6.8. Benefits and Limitation

The experimental results show the capability of our approach to get a performance that is similar to the best performing state-of-the-art approach used in this field (i.e., *Random Forests*), although it exploits only a class of data (i.e., the previous *legitimate* transactions), operating proactively.

Table 4: Performance

Dataset	Approach	F-score	AUC
ET	**DWT**	− 0.01	− 0.20
ET	**DFT**	− 0.06	− 0.21
GC	**DWT**	+ 0.08	− 0.27
GC	**DFT**	− 0.16	− 0.28

According to the obtained results, a benefit related to the adoption of our *fraud detection* approach is its capability to operate with a high level of data imbalance, a common scenario that leads towards a reduction of the effectiveness of the state-of-the-art solutions.

Considering that the proposed approach uses only a class of data to define its evaluation model, it operates in a proactive way, reducing/overcoming the *cold-start* problem.

These advantages can be exploited to define hybrid *fraud detection* approaches, which are able to work in different scenarios, blending the capabilities of our proactive approach with those of the non-proactive state-of-the-art ones.

We identify as the main limitation of our approach its application in those data scenarios characterized by a balanced distribution of data, although this can be considered a highly improbable scenario in the fraud detection context.

7. Conclusions and Future Work

Today, the sustainability represents an imperative paradigm to preserve the planet's resources. In this context, the new technologies allow us, in a more or less direct way, to adopt this paradigm in many day-to-day choices. The research that stand behind the *Big Data Analytics for Sustainability* is a representative example of such a scenario, since it aims to offer solutions able to allow people to exploit the new technologies in a smarter and secure way.

The exponential growth of the E-commerce market offers us the opportunity to buy or sell by adopting a large number of criteria such as, for instance, the sustainability one. However, this scenario is jeopardized by the risks related to the fraudulent use of electronic payment instruments, which represent the most common payment methods in such a environment.

For the aforementioned reason, the research that revolves around the *Big Data Information Security*, in this case that aimed to define effective fraud detection systems, assumes an increasingly central role, involving large investments by public and private entities. The risk scenario under consideration is mainly given by the combination of two factors: the exponential growth in the use of the E-commerce environment and the exponential growth in the use of credit cards by people. Considering that, as a result of these two factors, even money losses have reported an exponential trend in recent years, through this paper we wanted to investigate the benefits given by the adoption of proactive strategies of fraud detection.

The proactive approach proposed in this paper is able to face the problems discussed in Section 2.1, since its two strategies of implementation (i.e., *Fourier-based* and *Wavelet-based*) do not involve any previous *fraudulent* transaction, facing in this way the *data scarcity* and *data unbalance* problems. In addition, the adoption of a *transform-domain-based* model leads towards a better characterization of the class of transactions taken into account during the model definition (i.e., the *legitimate* one), reducing the problems related to the data variation (i.e., *non-adaptability* and *data heterogeneity* problems). Last but not least advantage of the proposed approach is related to its capability to face the *cold-start* problem.

Our results are interesting, considering that our state-of-the-art competitor (i.e., *Random Forests*) uses both classes of data (i.e., *legitimate* and *fraudulent*) to define its evaluation model, exploiting also a data balancing technique (i.e., *SMOTE*).

It should be observed that, more than wanting to replace the existing retroactive state-of-the-art approaches, through the proposed approach we want to introduce a novel proactive strategy that allows a fraud detection system to operate also in absence of previous fraudulent transactions. Such a capability can be exploited in a stand-alone fraud detection system or in order to define hybrid fraud detection solutions that combine the state-of-the-art retroactive approaches and our proactive approach.

The proposed proactive approach can be considered a valuable contribution in several *BDAS* research fields, such as that of the *Big Data Information Security for Sustainability* previously mentioned, or that of the *Computational intelligence and algorithms for Sustainability*, since it improves the state-of-the-art solutions, providing them with the capability to define an evaluation model on the basis of a single class of data.

For the aforementioned considerations, a possible future work could be focused on the definition of a novel fraud detection approach that combines the characteristics of the canonical non-proactive state-of-the-art approaches with those of our proactive approach, in order to define a hybrid strategy that maximizes the performance of both the approaches. Another interesting future work could be aimed to evaluate the advantages and disadvantages of our approach in scenarios that involve different kind of financial transaction data (e.g., those related to an E-commerce environment).

Acknowledgments

This research is partially funded by *Regione Sardegna* under project *Next generation Open Mobile Apps Development* (*NOMAD*), *Pacchetti Integrati di Agevolazione* (*PIA*) *Industria Artigianato e Servizi* (2013).

References

[1] V. Chang, Towards data analysis for weather cloud computing, Knowl.-Based Syst. 127 (2017) 29–45. `doi:10.1016/j.knosys.2017.03.003`.
URL `https://doi.org/10.1016/j.knosys.2017.03.003`

[2] M. V. M. Cano, F. Terroso-Saenz, A. González-Vidal, M. Valdés-Vela, A. F. Skarmeta, M. A. Zamora, V. Chang, Applicability of big data techniques to smart cities deployments, IEEE Trans. Industrial Informatics 13 (2) (2017) 800–809. doi:10.1109/TII.2016.2605581.
URL https://doi.org/10.1109/TII.2016.2605581

[3] T. Ferreira, I. Pedrosa, J. Bernardino, Business intelligence for e-commerce: Survey and research directions, in: Á. Rocha, A. M. R. Correia, H. Adeli, L. P. Reis, S. Costanzo (Eds.), Recent Advances in Information Systems and Technologies - Volume 1 [WorldCIST'17, Porto Santo Island, Madeira, Portugal, April 11-13, 2017]., Vol. 569 of Advances in Intelligent Systems and Computing, Springer, 2017, pp. 215–225. doi:10.1007/978-3-319-56535-4_22.
URL https://doi.org/10.1007/978-3-319-56535-4_22

[4] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.

[5] M. Lek, B. Anandarajah, N. Cerpa, R. Jamieson, Data mining prototype for detecting e-commerce fraud, in: S. Smithson, J. Gricar, M. Podlogar, S. Avgerinou (Eds.), Proceedings of the 9th European Conference on Information Systems, Global Co-operation in the New Millennium, ECIS 2001, Bled, Slovenia, June 27-29, 2001, 2001, pp. 160–165.

[6] A. J. Hoffman, R. E. Tessendorf, Artificial intelligence based fraud agent to identify supply chain irregularities, in: M. H. Hamza (Ed.), IASTED International Conference on Artificial Intelligence and Applications, part of the 23rd Multi-Conference on Applied Informatics, Innsbruck, Austria, February 14-16, 2005, IASTED/ACTA Press, 2005, pp. 743–750.

[7] K. M. Gopinathan, L. S. Biafore, W. M. Ferguson, M. A. Lazarus, A. K. Pathria, A. Jost, Fraud detection using predictive modeling, uS Patent 5,819,226 (Oct. 6 1998).

[8] M. E. Edge, P. R. F. Sampaio, A survey of signature based methods for financial fraud detection, Computers & Security 28 (6) (2009) 381–394. doi:10.1016/j.cose.2009.02.001.
URL https://doi.org/10.1016/j.cose.2009.02.001

[9] M. J. Lenard, P. Alam, Application of fuzzy logic fraud detection, in: M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Technology (5 Volumes), Idea Group, 2005, pp. 135–139.

[10] Y. Sahin, S. Bulkan, E. Duman, A cost-sensitive decision tree approach for fraud detection, Expert Syst. Appl. 40 (15) (2013) 5916–5923. doi:10.1016/j.eswa.2013.05.021.
URL https://doi.org/10.1016/j.eswa.2013.05.021

[11] D. G. Whiting, J. V. Hansen, J. B. McDonald, C. C. Albrecht, W. S. Albrecht, Machine learning methods for detecting patterns of management fraud, Computational Intelligence 28 (4) (2012) 505–527.

[12] L. Zhang, J. Yang, W. Chu, B. L. Tseng, A machine-learned proactive moderation system for auction fraud detection, in: C. Macdonald, I. Ounis, I. Ruthven (Eds.), Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, ACM, 2011, pp. 2501–2504. doi:10.1145/2063576.2064002.
URL http://doi.acm.org/10.1145/2063576.2064002

[13] C. Assis, A. M. Pereira, M. de Arruda Pereira, E. G. Carrano, Using genetic programming to detect fraud in electronic transactions, in: C. V. S. Prazeres, P. N. M. Sampaio, A. Santanchè, C. A. S. Santos, R. Goularte (Eds.), A Comprehensive Survey of Data Mining-based Fraud Detection Research, Vol. abs/1009.6119, 2010, pp. 337–340.

[14] B. Hooi, N. Shah, A. Beutel, S. Günnemann, L. Akoglu, M. Kumar, D. Makhija, C. Faloutsos, BIRDNEST: bayesian inference for ratings-fraud detection, in: S. C. Venkatasubramanian, W. M. Jr. (Eds.), Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016, SIAM, 2016, pp. 495–503. doi:10.1137/1.9781611974348.56.
URL https://doi.org/10.1137/1.9781611974348.56

[15] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32. doi:10.1023/A:1010933404324.

[16] S. K. Katsikas, J. López, G. Pernul, Trust, privacy and security in e-business: Requirements and solutions, in: Panhellenic Conference on Informatics, Vol. 3746 of Lecture Notes in Computer Science, Springer, 2005, pp. 548–558.

[17] M. E. Edge, P. R. F. Sampaio, M. Choudhary, Towards a proactive fraud management framework for financial data streams, in: Third IEEE International Symposium on Dependable, Autonomic and Secure Computing, DASC 2007, Columbia, MD, USA, September 25-26, 2007, IEEE Computer Society, 2007, pp. 55–64. doi:10.1109/DASC.2007.25.
URL https://doi.org/10.1109/DASC.2007.25

[18] R. Saia, S. Carta, Evaluating credit card transactions in the frequency domain for a proactive fraud detection approach, in: SECRYPT, SciTePress, 2017, pp. 335–342.

[19] R. Saia, A discrete wavelet transform approach to fraud detection, in: NSS, Vol. 10394 of Lecture Notes in Computer Science, Springer, 2017, pp. 464–474.

[20] P. Duhamel, M. Vetterli, Fast fourier transforms: a tutorial review and a state of the art, Signal processing 19 (4) (1990) 259–299.

[21] P. Chaovalit, A. Gangopadhyay, G. Karabatis, Z. Chen, Discrete wavelet transform-based time series analysis and mining, ACM Comput. Surv. 43 (2) (2011) 6:1–6:37. doi:10.1145/1883612.1883613.
URL http://doi.acm.org/10.1145/1883612.1883613

[22] E. J. Keogh, A decade of progress in indexing and mining large time series databases, in: U. Dayal, K. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, Y. Kim (Eds.), Proceedings of the 32nd International Conference on Very Large Data Bases, Seoul, Korea, September 12-15, 2006, ACM, 2006, p. 1268.
URL `http://dl.acm.org/citation.cfm?id=1164262`

[23] E. Baidoo, J. L. Priestley, An analysis of accuracy using logistic regression and time series.

[24] K. R. Lai, C. Fan, W. Huang, P. Chang, Evolving and clustering fuzzy decision tree for financial time series data forecasting, Expert Syst. Appl. 36 (2) (2009) 3761–3773. `doi:10.1016/j.eswa.2008.02.025`.
URL `https://doi.org/10.1016/j.eswa.2008.02.025`

[25] R. Saia, S. Carta, A frequency-domain-based pattern mining for credit card fraud detection, in: M. Ramachandran, V. M. Muñoz, V. Kantere, G. Wills, R. J. Walters, V. Chang (Eds.), Proceedings of the 2nd International Conference on Internet of Things, Big Data and Security, IoTBDS 2017, Porto, Portugal, April 24-26, 2017, SciTePress, 2017, pp. 386–391. `doi:10.5220/0006361403860391`.
URL `https://doi.org/10.5220/0006361403860391`

[26] D. Zheng, F. Li, T. Zhao, Self-adaptive statistical process control for anomaly detection in time series, Expert Syst. Appl. 57 (2016) 324–336. `doi:10.1016/j.eswa.2016.03.029`.
URL `https://doi.org/10.1016/j.eswa.2016.03.029`

[27] K. A. Salleh, L. Janczewski, Technological, organizational and environmental security and privacy issues of big dat
Procedia Computer Science 100 (2016) 19 – 28, international Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016. `doi:http://dx.doi.org/10.1016/j.procs.2016.09.119`.
URL `http://www.sciencedirect.com/science/article/pii/S1877050916322864`

[28] N. G. Miloslavskaya, A. Makhmudova, Survey of big data information security, in: M. Younas, I. Awan, J. E. Haddad (Eds.), 4th IEEE International Conference on Future Internet of Things and Cloud Workshops, FiCloud Workshops 2016, Vienna, Austria, August 22-24, 2016, IEEE Computer Society, 2016, pp. 133–138. `doi:10.1109/W-FiCloud.2016.38`.
URL `https://doi.org/10.1109/W-FiCloud.2016.38`

[29] S. Kokolakis, Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenom
Computers & Security 64 (2017) 122–134. `doi:10.1016/j.cose.2015.07.002`.
URL `https://doi.org/10.1016/j.cose.2015.07.002`

[30] R. Saia, L. Boratto, S. Carta, Multiple behavioral models: A divide and conquer strategy to fraud detection in financ
in: A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, J. Filipe (Eds.), KDIR

2015 - Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Volume 1, Lisbon, Portugal, November 12-14, 2015, SciTePress, 2015, pp. 496–503. doi:10.5220/0005637104960503.
URL https://doi.org/10.5220/0005637104960503

[31] M. Ahmed, A. N. Mahmood, M. R. Islam, A survey of anomaly detection techniques in financial domain, Future Generation Computer Systems 55 (2016) 278–288.

[32] S. Sorournejad, Z. Zojaji, R. E. Atani, A. H. Monadjemi, A survey of credit card fraud detection techniques: Data and technique oriented perspective, CoRR abs/1611.06439.
URL http://arxiv.org/abs/1611.06439

[33] A. Chatterjee, A. Segev, Data manipulation in heterogeneous databases, ACM SIGMOD Record 20 (4) (1991) 64–68.

[34] D. Che, M. S. Safran, Z. Peng, From big data to big data mining: Challenges, issues, and opportunities, in: B. Hong, X. Meng, L. Chen, W. Winiwarter, W. Song (Eds.), Database Systems for Advanced Applications - 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, April 22-25, 2013. Proceedings, Vol. 7827 of Lecture Notes in Computer Science, Springer, 2013, pp. 1–15. doi:10.1007/978-3-642-40270-8_1.
URL https://doi.org/10.1007/978-3-642-40270-8_1

[35] I. Brown, C. Mues, An experimental comparison of classification algorithms for imbalanced credit scoring data sets, Expert Syst. Appl. 39 (3) (2012) 3446–3453. doi:10.1016/j.eswa.2011.09.033.

[36] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284. doi:10.1109/TKDE.2008.239.

[37] V. Vinciotti, D. J. Hand, Scorecard construction with unbalanced class sizes, Journal of Iranian Statistical Society 2 (2) (2003) 189–205.

[38] A. I. Marqués, V. García, J. S. Sánchez, On the suitability of resampling techniques for the class imbalance problem in credit scoring, JORS 64 (7) (2013) 1060–1070. doi:10.1057/jors.2012.120.
URL http://dx.doi.org/10.1057/jors.2012.120

[39] S. F. Crone, S. Finlay, Instance sampling in credit scoring: An empirical study of sample size and balancing, International Journal of Forecasting 28 (1) (2012) 224–238.

[40] O. Loyola-González, J. F. Martínez Trinidad, J. A. Carrasco-Ochoa, M. García-Borroto, Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases,

Neurocomputing 175 (2016) 935–947. doi:10.1016/j.neucom.2015.04.120.
URL https://doi.org/10.1016/j.neucom.2015.04.120

[41] J. Attenberg, F. J. Provost, Inactive learning?: difficulties employing active learning in practice, SIGKDD Explorations 12 (2) (2010) 36–41. doi:10.1145/1964897.1964906.
URL http://doi.acm.org/10.1145/1964897.1964906

[42] J. T. Wells, Corporate fraud handbook: Prevention and detection, John Wiley & Sons, 2017.

[43] Z. Rezaee, Financial statement fraud: prevention and detection, John Wiley & Sons, 2002.

[44] C. Phua, V. C. S. Lee, K. Smith-Miles, R. W. Gayler, A comprehensive survey of data mining-based fraud detection research, CoRR abs/1009.6119.

[45] A. Abdallah, M. A. Maarof, A. Zainal, Fraud detection system: A survey, Journal of Network and Computer Applications 68 (2016) 90–113.

[46] M. R. Chernick, Wavelet methods for time series analysis, Technometrics 43 (4) (2001) 491. doi:10.1198/tech.2001.s49.
URL http://dx.doi.org/10.1198/tech.2001.s49

[47] D. B. Percival, A. T. Walden, Wavelet methods for time series analysis, Vol. 4, Cambridge university press, 2006.

[48] S. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693. doi:10.1109/34.192463.
URL http://dx.doi.org/10.1109/34.192463

[49] P. M. Higgins, Professor Higgins's Problem Collection, Oxford University Press, 2017.

[50] S. W. Smith, et al., The scientist and engineer's guide to digital signal processing.

[51] S. Lessmann, B. Baesens, H. Seow, L. C. Thomas, Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, European Journal of Operational Research 247 (1) (2015) 124–136. doi:10.1016/j.ejor.2015.05.030.

[52] S. Bhattacharyya, S. Jha, K. K. Tharakunnel, J. C. Westland, Data mining for credit card fraud: A comparative study, Decision Support Systems 50 (3) (2011) 602–613. doi:10.1016/j.dss.2010.08.008.
URL http://dx.doi.org/10.1016/j.dss.2010.08.008

[53] A. L. Buczak, E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detect
IEEE Communications Surveys and Tutorials 18 (2) (2016) 1153–1176. doi:10.1109/COMST.2015.2494502.
URL https://doi.org/10.1109/COMST.2015.2494502

[54] J. T. S. Quah, M. Sriganesh, Real-time credit card fraud detection using computational intelligence, Expert Syst. Appl. 35 (4) (2008) 1721–1732. doi:10.1016/j.eswa.2007.08.093.
URL http://dx.doi.org/10.1016/j.eswa.2007.08.093

[55] J. Dean, S. Ghemawat, Mapreduce: simplified data processing on large clusters, Commun. ACM 51 (1) (2008) 107–113. doi:10.1145/1327452.1327492.
URL http://doi.acm.org/10.1145/1327452.1327492

[56] A. D. Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, Calibrating probability with undersampling for unbalanced classification, in: IEEE Symposium Series on Computational Intelligence, SSCI 2015, Cape Town, South Africa, December 7-10, 2015, IEEE, 2015, pp. 159–166. doi:10.1109/SSCI.2015.33.
URL http://dx.doi.org/10.1109/SSCI.2015.33

[57] D. Faraggi, B. Reiser, Estimation of the area under the roc curve, Statistics in medicine 21 (20) (2002) 3093–3106.

[58] K. W. Bowyer, N. V. Chawla, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, CoRR abs/1106.1813.
URL http://arxiv.org/abs/1106.1813