ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal
Regular Issue, Vol. 10 N. 1 (2021), 63-76
eISSN: 2255-2863
DOI: https://doi.org/10.14201/ADCAIJ20211016376

# Explainable Credit Card Fraud Detection with Image Conversion

Duygu Sinanç Terzi[a], Umut Demirezen[b] and Şeref Sağıroğlu[a]

[a] Department of Computer Engineering, Faculty of Engineering, Gazi University, Ankara, Turkey

[b] Artificial Intelligence and Big Data Unit, Digital Transformation Office, Presidency of the Republic of Turkey, Ankara, Turkey

duygusinanc@gazi.edu.tr, umut@demirezen.tech, ss@gazi.edu.tr

KEYWORD

*Fraud detection; Time series; Deep learning; Explainable artificial intelligence; Image conversion*

ABSTRACT

*The increase in the volume and velocity of credit card transactions causes class imbalance and concept deviation problems in data sets where credit card fraud is detected. These problems make it very difficult for traditional approaches to produce robust detection models. In this study, a different perspective has been developed for this problem and a novel approach named Fraud Detection with Image Conversion (FDIC) is proposed. FDIC handles credit card transactions as time series and transforms them into images. These images, which comprise temporal correlations and bilateral relationships of features, are classified by a convolutional neural network architecture as fraudulent or legitimate. When the obtained results are compared with the related studies, FDIC has the best F1-score and recall values, which are 85.49% and 80.35%, respectively. This shows that FDIC is better than other studies in detecting fraudulent instances associated with high cost. Since the images created during the FDIC process are difficult to interpret, a new explainable artificial intelligence approach is also presented. In this way, feature relationships that have a dominant effect on fraud detection are revealed.*

## 1. Introduction

Technological and financial innovations in the payment industry offer more advanced payment methods and affect consumers' choice of payment instruments. Payment cards are becoming the premier payment method for in-person and online purchases. As the spending behavior of consumers has improved, the global market for payment cards has grown substantially. On a global scale, credit cards,

debit cards and prepaid cards generated 43.916 trillion dollars transaction volume in purchases of goods and services and cash advances and withdrawals in 2019 (Nilson, 2020). Innovations in payment methods provide greater consumer comfort and efficiency while the complexity of new technologies has led to the emergence of new risk factors. Although it is generally known that less than 0.1% of banking transactions are fraudulent (Gold, 2014), the financial loss caused by this ratio is quite high. The frauds committed offline by somehow seizing the card or online with/without a card caused 30.07 billion dollars globally in 2019, and this figure is estimated to reach 35.67 billion dollars in 2023 (Nilson, 2020). Considering that around 5% of potentially relevant data was tagged (Gupta et al., 2019), the difficulty of analytics to prevent losses caused by fraud also increases significantly.

Credit card fraud activities occur when fraudsters exploit credit cards for personal interests without the knowledge of the cardholder and the card provider. These activities are usually carried out by purchases, withdrawing money from the automated teller machine, and transactions from a workplace, online platforms, or telephone banking. In addition to putting serious financial burdens on cardholders and card providers, actions taken in this context have wider impacts, such as funding for illegal activities.

In the process of combating fraud, prevention and detection mechanisms are used to block or reduce frauds and catch new threats (Behdad et al., 2012). Prevention is the first layer that protects systems against fraud, and it restricts, suppresses, destroys, controls, or removes the occurrence of abuses. Examples of prevention mechanisms are actions such as encryption of data, use of firewalls, or active inspection of network flow. Detection is the next layer of protection. Fraud detection aims to discover and identify system intrusions and fraudulent activities and report them to the system administrator. Although fraud was initially attempted to be discovered using manual inspection techniques, these complex and time-consuming techniques have been replaced by automated fraud detection systems. When fraudulent activities are detected and the type of fraud is verified, vulnerabilities are investigated in detail and precautions are taken.

Fraud detection systems tend to fail as they involve challenging technical and methodological problems, have low accuracy rates, and generate many false alarms (Arif et al., 2015). Fraud detection approaches in the literature can be grouped as anomaly detection, misuse detection, and hybrid detection (Adewumi et al., 2017). Anomaly detection approaches are based on behavioral profiling methods in which the behavior of each individual is modeled and deviations are monitored. Misuse detection approaches use rule-based, statistical, or heuristic methods to detect suspicious transactions in line with known fraudulent behaviors. Anomaly detection suffers from a lack of generalization capability and the presence of high false alarm rates while misuse detection cannot detect new types of fraud. For this reason, hybrid detection includes both anomaly and misuse approaches.

There is always a strong periodic pattern in spending behavior for every cardholder, and anomalies in transactions can be detected by analyzing past transactions of the individual cardholder. Based on this fact, fraud detection can also be treated as a time series problem (Seyedhossein et al., 2010).

Time series represents the collection of data obtained from sequential measurements over time. Due to their unique structure, time series poses challenges for classic data mining tasks. Time series mining, which is a customized type of data mining, aims to identify models that best describe the underlying causes of the observed time series components. Time series mining is carried out for seven purposes: classification, indexing, clustering, forecasting, summarization, anomaly detection, and segmentation (Maimon et al., 2005).

While performing analyses for these purposes, the problem arises of representing the silent features that help identify important parts of the time series. The problem of time series data representation

becomes increasingly complex due to the high dimensionality of time series, presence of excessive noise, non-linear relationship of data elements, and frequent updating nature of data (Wilson, 2017). Therefore, any data representation method needs to be robust against random noise and bring the data to a manageable size while preserving the important characteristics of the original data. Converting time series to another domain is generally performed with approaches such as sampling, approximation, symbolic data representations, or indexing (Fu, 2011).

By presenting a different perspective in time series data representation, the studies that transform time series into images have gained importance in recent years. These studies, which are used in many different application areas, were carried out for the following purposes: human activity recognition (Qin et al., 2020), day-ahead solar irradiation forecasting (Hong et al., 2020), traffic incident detection (Liu et al., 2020), single residential load forecasting (Estebsari et al., 2020), fault diagnosis of induction motors (Hsueh et al., 2020), and fraud detection (Zhang et al., 2018).

Therefore, in this paper, a new approach called Fraud Detection with Image Conversion (FDIC) has been proposed, which aims to both reveal hidden patterns by transferring the characteristics of the time series to two-dimensional images and overcome drawbacks of traditional fraud detection approaches. According to FDIC, credit card transactions were considered as time series, these transactions were transformed into images, and these images were classified as fraudulent or legitimate. After determining the most appropriate technique that converts the time series into images for fraud detection, the obtained FDIC results were compared with related studies. Since the created images are difficult to interpret, the results of the FDIC were also evaluated in terms of explainable artificial intelligence to provide a window into the data and prediction. In this way, regions/features where FDIC focuses on both fraudulent and legitimate classes during the classification were visualized on a heat map.

The paper is organized as follows: in Section 2, the methodology of the proposed FDIC approach is explained. In Section 3, the experimental studies, obtained results, and the explainability of the proposed model are presented. Finally, in Section 4, we conclude the paper with future direction and discussion.

## 2. Methodology

FDIC is mainly carried out in two stages: conversion of credit card transactions into images and classification of these images. In image conversion step, three different techniques are used: Gramian Angular Fields, Markov Transition Fields, and Recurrence Plot. In the classification step, images obtained with these techniques are classified with a Convolutional Neural Network architecture. The methodology of FDIC is summarized in Figure 1.

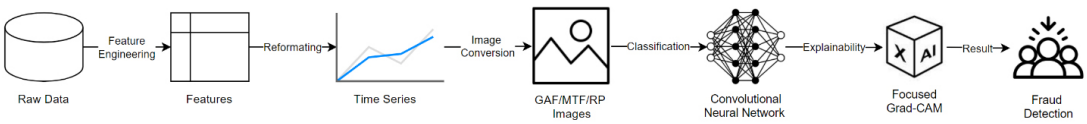



*Figure 1. The methodology of FDIC*

## 2.1. Conversion of Time Series into Images

In the process of converting time series to images, generally three techniques are used, which are Gramian Angular Fields (GAF), Markov Transition Fields (MTF), and Recurrence Plot (RP). GAF represents time series in polar coordinate system and generates images by expressing polar angles as asymmetry matrix. MTF represents a domain of transition probabilities for a discretized time series. Lastly, RP represents the distances between trajectories extracted from the original time series.

To reveal the differences and provide basic understanding of the methods, GAF, MTF, and RP images obtained from a sample sinusoidal signal are given in Figure 2.

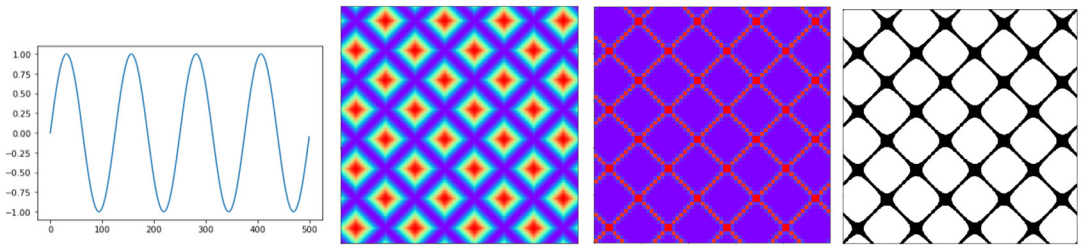



*Figure 2. GASF (b), MTF (c), and RP (d) images generated from a sample sinusoidal signal (a)*

A time series $X = \{x_1, \ldots x_n\}$ of size n is an ordered sequence of real-value data. When $X$ is rescaled in the interval [0,1] by Eq. 1, $\tilde{X}$ is represented in polar coordinates by encoding the value as angular cosine with Eq. 2, where r is radius, $\varnothing$ is angle polar coordinates, $t_i$, is the time stamp and $N$ is a constant factor for regularization. After the rescaling and transforming time series into the polar coordinate system, it is easily reconsidered by the trigonometric summation. Thus, Gramian Angular Summation Fields (GASF), a type of GAF, is obtained as given in Eq. 3 and Eq. 4, where I is a unit row vector (Wang et al., 2015a).

$$\tilde{x}_0^i = \frac{x_i - \min(X)}{\max(X) - \min(X)} \tag{1}$$

$$\begin{cases} \emptyset = \arccos(\tilde{x}_i), -1 \leq \tilde{x}_i \leq 1, \tilde{x}_i \in \tilde{X} \\ r = \frac{t_i}{N}, t_i \in \mathbb{N} \end{cases} \tag{2}$$

$$GASF = \begin{bmatrix} \cos(\emptyset_1 + \emptyset_1) & \cdots & \cos(\emptyset_1 + \emptyset_n) \\ \vdots & \ddots & \vdots \\ \cos(\emptyset_n + \emptyset_1) & \cdots & \cos(\emptyset_n + \emptyset_n) \end{bmatrix} \tag{3}$$

$$= \tilde{X}' . \tilde{X} - \sqrt{I - \tilde{X}^2}' . \sqrt{I - \tilde{X}^2} \tag{4}$$

Given a time series $X$, the $Q$ quantile bins can be defined and each $x_i$ can be assigned to corresponding bins $q_j$ . By counting transitions between quantile bins in the first order Markov chain along the time axis, a $QxQ$ weighted adjacency matrix $W$ is formed. After normalization, MTF is defined as given in Eq. 5 (Wang et al., 2015b).

$$MTF = \begin{bmatrix} w_{ij|x_1 \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_1 \in q_i, x_n \in q_j} \\ \vdots & \ddots & \vdots \\ w_{ij|x_n \in q_i, x_1 \in q_j} & \cdots & w_{ij|x_n \in q_i, x_n \in q_j} \end{bmatrix} \quad (5)$$

Given a time series *X*, the extracted trajectories can be defined as $\vec{x}_i$. The pairwise distance between the trajectories, *RP* is obtained as given in Eq. 6, where is a threshold and ∅ is Heaviside function as given in Eq. 7 (Hsueh et al., 2020). In brief, if the distance between two points is less than the specified threshold value, this point is defined as recurrence and expressed as a black pixel in the image.

$$R_{i,j} = \theta\left(\varepsilon - \left\|\vec{x}_i - \vec{x}_j\right\|\right) \quad (6)$$

$$\theta(x) = 0,\ if\ x < 0\ ve\ \theta(x) = 1,\ others) \quad (7)$$

Converting time series to images ensures both the preservation of temporal dependence and revealing of all bilateral relationships between variables. Thus, unlike classical methods that model the effect of the feature on the label, the positive or negative effect of relationships between features on the label will be modeled.

## 2.2. Classification

Time series classification is training a classifier on a dataset D = { ($X_1$, $Y_1$), ($X_2$, $Y_2$), ..., ($X_n$, $Y_n$)} consisting of *X,Y* pairs where *X* is a time series with *Y* as its corresponding class label to map from the space of possible inputs to a probability distribution over the labels. There has been a wide range of approaches to time series classification. These approaches can be divided into six groups: traditional methods analyzing the whole series, methods using one or more intervals of the series, methods that find short patterns called shapelets, dictionary-based methods obtained with the frequency counts of repeating patterns, model-based methods that fit a generative model to each series, and a combination of two or more methods (Bagnall et al, 2017). In recent years, deep learning solutions gain importance for time series classification due to their ability to deal with missing values and noise, adapt to multivariate inputs, and handle non-linear functions. Deep learning approaches for the time series classification are examined in two categories as generative and discriminative (Fawaz et al., 2019). While generative approaches aim to find a good representation of time series before the training phase, discriminative approaches learn directly from raw time series or hand-designed features as a result of various transformations.

Convolutional Neural Network (CNN) is one of the most widely preferred deep learning model in diverse computer vision applications where multiple layers are trained quite effectively. A basic CNN generally consists of three main neural layer types: convolutional layer, pooling/subsampling layer, and fully connected layer (Gu et al., 2018). Convolutional layer is comprised of convolution kernels which are used to compute different feature maps. Pooling layer is used to reduce the dimensions of feature maps and network parameters. Finally, fully connected layer acts like a traditional neural network and enables classification, regression, or input into other network based on the features extracted by the previous layers.

$$X_k^l = \sigma\left(W_k^{l-1} * X^{l-1} + b_k^{l-1}\right) \quad (8)$$

At each layer, the input image is convolved with a set of $K$ kernels $w = \{W_1, ... W_k\}$ and added biases $B = \{b_1, ... b_k\}$ each generating a new feature map $X_k$. These features are subjected to an element-wise non-linear transform and the same process is repeated for every convolutional layer $l$ as given in Eq. 8 (Ma et al., 2019). In the simplest case, these layers transform an image volume into an output volume.

# 3. Experimental Study

In this section, the dataset used in the experiments is explained, the CNN architecture used for classification is detailed, the experimental results are presented comparatively, and the obtained results are evaluated in terms of explainable artificial intelligence.

## 3.1. Dataset

The dataset used in analysis includes transactions that are built and shared by European cardholders with credit cards in two days (ULB, 2020). Due to confidentiality issues, features V1, V2,…, V28 were shared as the result of Principal Component Analysis (PCA). The dataset consists of 284807 legitimate and 492 fraudulent transactions. After the duplicate records are removed, transactions are dropped to 283253 legitimate and 473 fraudulent.

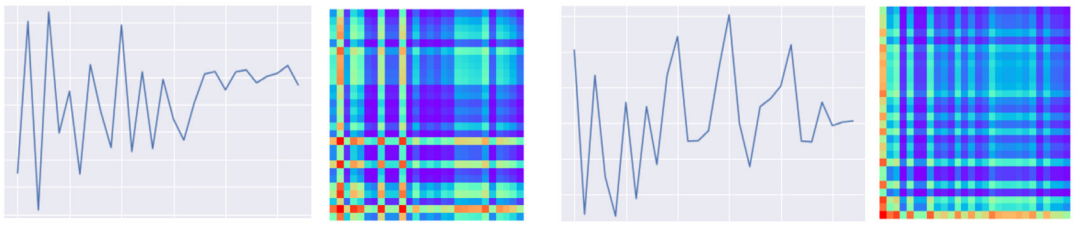



*Figure 3. Original signal (a,c) and GASF image (b,d) of data known as fraudulent and legitimate, respectively*

The ability to gain great insights from reliable, contextualized and consumable data at any scale is possible by transforming the data into smart data. For this purpose, credit card transactions that will be input for classification with CNN are encoded as images. Since there are 28 variables as a result of PCA, the produced image dimensions are 28x28. After the time series characteristics are transferred to the two-dimensional images from different features such as colors, lines, or points as seen an example in Fig. 3, the classification process is initiated.

## 3.2. Convolutional Neural Network Architecture

When designing the CNN architecture, the aim is not to construct complex structures that produce quite high success, but to demonstrate the proof of concept. In addition, the small size of the input images made it impossible to design complex architectures. Therefore, the proposed CNN consists of five convolutional layers, five batch normalization layers, and three pooling layers as shown in Figure 4.

Each convolutional layer has 2x2 kernel sizes. Exponential linear unit (ELU) is applied to all convolutional layers as activation function. ELU, which is given in Eq. 9, is in tendency to converge cost to zero faster by allowing negative input (Clevert, et al., 2015). In the last layer, softmax function, which calculates the probabilities of each target class on all possible target classes, is used as given in Eq. 10.

$$f(x) = \begin{cases} x, & x > 0 \\ \alpha(e^x - 1), & x \leq 0 \end{cases} \tag{9}$$

$$\sigma(x)_i = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_i}} \tag{10}$$

As pooling layer takes the convolution layer as input, max pooling with 2x2 kernel size is used. After reshaping the tensor in flatten layer, a dropout with 0.1 rate is applied to prevent overfitting. ELU function is used in the first three fully connected layers, and softmax function is used in the last fully connected layer. Lastly, the loss function is categorical cross entropy, which is optimized using adaptive moment estimation (ADAM). ADAM maintains adaptive learning rates for each parameter, where $g_t$ is gradient, $m_t$ is exponential average of gradient, $v_t$ is exponential average of squares of gradient, $\beta_1$ and $\beta_2$ are hyperparameters, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected estimates, and $\varnothing_t$ is parameter to be updated (Kingma et al., 2014).

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{11}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{12}$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \tag{13}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{14}$$

$$\theta_t = \theta_{t-1} - \frac{\alpha \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{15}$$

To measure the performance of the classification model, categorical cross entropy is used as a loss function. In our case, loss function evaluates a candidate solution for fraudulent or legitimate transactions. Given a true label $y$ and a predicted label $p$, it is defined as given in Eq. 16 for binary classes (Rusiecki, 2019).

$$L(y, p) = -\sum_{i=1}^{2} y_i \log p_i \tag{16}$$

In the next step, classification performance is evaluated through various metrics. Fraud detection is naturally a class imbalanced data problem in which the fraudulent class is low and the legitimate class is too much in the data set. When it comes to class imbalance data, accuracy and AUC appear to be inappropriate in most real domains because they have bias towards the majority class examples. $F_1$-score is a commonly used metric to evaluate the performance of imbalanced classification models because it provides a way to combine both precision and recall into a single measure (Guo et al., 2016). For this reason, in the evaluation of the results of this study, more importance was attached to the success in the $F_1$-score.

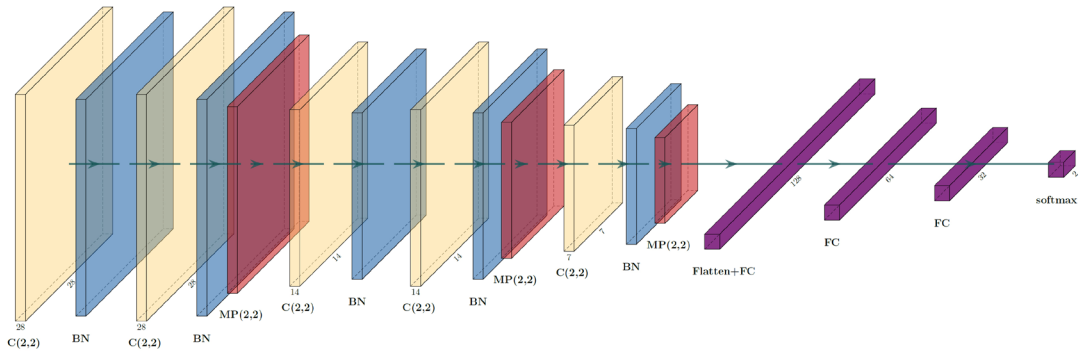



*Figure 4. The proposed CNN architecture*

To evaluate the performance of FDIC, 5-folds cross validation is utilized. To create unbiased folds, the splitting process is performed with the stratification technique. In this way, all folds have the same imbalance ratio. Each fold is in turn used for the test while the other four folds are used for training in classification. Inherently, 80% of the original dataset is used for training and remaining 20% is used for test. The CNN architecture is trained for 100 epochs. After 100 epochs, the model with the best $F_1$-score is saved and this model is applied to the test set. In accordance with the metrics obtained in this way, experimental results are evaluated.

## 3.3. Experimental Results

Experiments were carried out on NVIDIA TESLA V100 Tensor Core GPU with 32GB. Two experiments were organized to evaluate the success of FDIC: determining the most suitable conversion technique for fraud detection and performance comparison with related studies using the same dataset.

In the first experiment, the most suitable technique that converts the time series into images for fraud detection was determined. Images acquired with GASF, MTF, and RP were classified using the proposed CNN architecture. As given in Table 1, the best result in $F_1$-score was obtained when GASF images were used.

*Table 1. Classification results of different image transformation techniques*

| Conversion of Time Series Into Images | ACC | Precision | Recall | F1-score |
|---|---|---|---|---|
| GASF | 0,9960 | **0,9294** | **0,8404** | **0,8827** |
| MTF | 0,9988 | 0,8181 | 0,3829 | 0,5217 |
| RP | **0,9992** | 0,8472 | 0,6489 | 0,7349 |

After GASF images were found to be more effective in detecting fraud, the second experiment was conducted to evaluate the obtained results. FDIC was compared with other related studies that used the same data set. As given in Table 2, after 5-folds cross validation, FDIC produced the highest $F_1$-score. It seems that the FDIC creates a good balance between recall, which gives the success in classifying fraudulent instances, and precision, which gives the success in cases classified as fraud. This shows that the FDIC is better than other studies in detecting fraudulent instances associated with high cost.

*Table 2. Comparison with related studies*

| Reference | Approach | ACC | Precision | Recall | $F_1$-score |
|---|---|---|---|---|---|
| Abakarim, et al., 2018 | ROS+Linear SVM Regression | 0,9801 | 0,1391 | 0,593 | 0,2253 |
| Abakarim, et al., 2018 | ROS+Logistic Regression | 0,9868 | 0,1897 | 0,528 | 0,2791 |
| Abakarim, et al., 2018 | ROS+NN Based Classification | 0,9794 | 0,1470 | 0,6598 | 0,2404 |
| Abakarim, et al., 2018 | ROS+Non-Linear Auto-regression | 0,9659 | 0,1000 | 0,7401 | 0,1762 |
| Abakarim, et al., 2018 | ROS+Deep NN Autoencoders | 0,9855 | 0,1972 | 0,5821 | 0,2947 |
| Xuan et al., 2018 | GAN+DNN (N=630) | **0,9996** | 0,9320 | 0,7328 | 0,8205 |
| Xuan et al., 2018 | SMOTE+DNN (N=630) | **0,9996** | **0,9680** | 0,6946 | 0,8088 |
| Al-Shabi, 2019 | Logistic Regression | 0,9991 | 0,9300 | 0,5700 | 0,7100 |
| Al-Shabi, 2019 | Autoencoder (Threshold=5) | 0,9870 | 0,0110 | 0,6400 | 0,1900 |
| Mohammed et al., 2018 | SMOTE+Balanced Bagging Ensemble | * | 0,9412 | 0,7273 | 0,8205 |
| Mohammed et al., 2018 | SMOTE+RF | * | 0,4324 | 0,7273 | 0,5424 |
| Mohammed et al., 2018 | SMOTE+Gaussian Naive Bayes | * | 0,0786 | 0,5000 | 0,1358 |
| Proposed Approach | FDIC | 0,9995 | 0,9146 | **0,8035** | **0,8549** |

*: Unspecified

It has been observed that there are various differences in the application details of the related studies. In the FDIC training process, contrary to (Abakarim, et al., 2018) and (Al-Shabi, 2019), duplicate records were removed to avoid bias. Unlike studies aimed at increasing minority class samples by over-sampling (Abakarim, et al., 2018) or creating synthetic data (Xuan et al., 2018; Mohammed et al., 2018) to deal with class imbalance, FDIC did not perform any sampling on the original data. Another difference is that, Mohammed et al. (2018) divided the dataset as 90% for training and 10% for test. This may lead to over-fitting and misleading results may occur.

FDIC reveals quite promising results for fraud detection by treating a single image as an input feature and modeling the relationships between features, compared to methods that require more than one feature and model the relationship of each feature on the target variable. This situation shows that moving the features to a different relation space can increase the model performance.

## 3.4. Explainability

Even if deep learning models achieve impressive prediction success, they are typically considered as black boxes. In recent years, the techniques aimed to open black box models have helped a better understanding of what the model has learned. In this context, the explainable artificial intelligence

approaches provide such benefits as: to justify the decisions as being ethical and fair, to control decisions to keep things from going wrong, to improve the model, and to discover new patterns (Adadi et al., 2018).

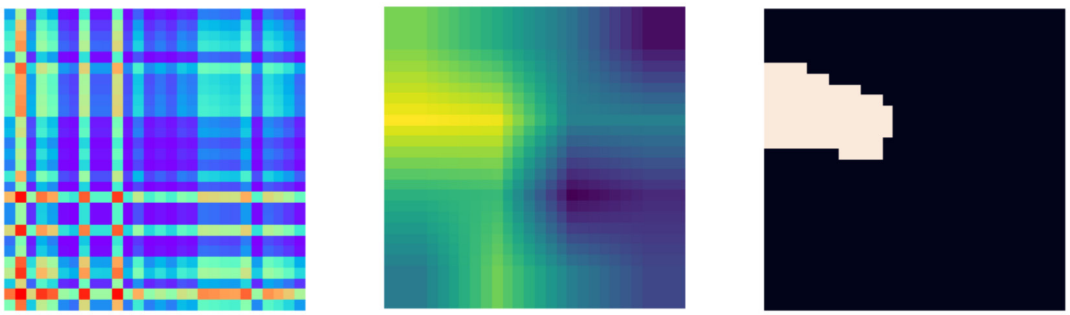
The literature on explainable artificial intelligence methods is examined in three classes: global, local, and introspective (Guidotti et al., 2018). Global methods intend to provide an overall approximation of the behavior of the black-box models. Local methods reveal the decisions made by the black-box model over specific results or samples of a dataset. Lastly, introspective methods generate the explanations by relating inputs to outputs of a black-box model.

Gradient-weighted Class Activation Mapping (Grad-CAM) is a local method that generates visual explanations to make CNN based models more transparent (Selvaraju et al., 2017). Grad-CAM produces a localization map $L$ which highlights the regions of the input image that are mostly influenced by the classifier. To obtain $L$, firstly the gradient of the score for class $c$, $y_c$ is computed by feature maps $A^k$ of a convolution layer. These gradients are pooled to get weights $\alpha_k^c$. Then, Grad-CAM is procured as a weighted combination of feature maps and introduced as follows.

$$y_c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A_{ij}^k \tag{17}$$

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{18}$$

$$L_{Grad-CAM}^c = \text{RELU}(\sum_k \alpha_k^c A^k) \tag{19}$$



*(a)* *(b)* *(c)*

*Fig. 5. For a fraudulent data GASF (a), Grad-CAM (b) and Focused Grad-CAM (c) images*

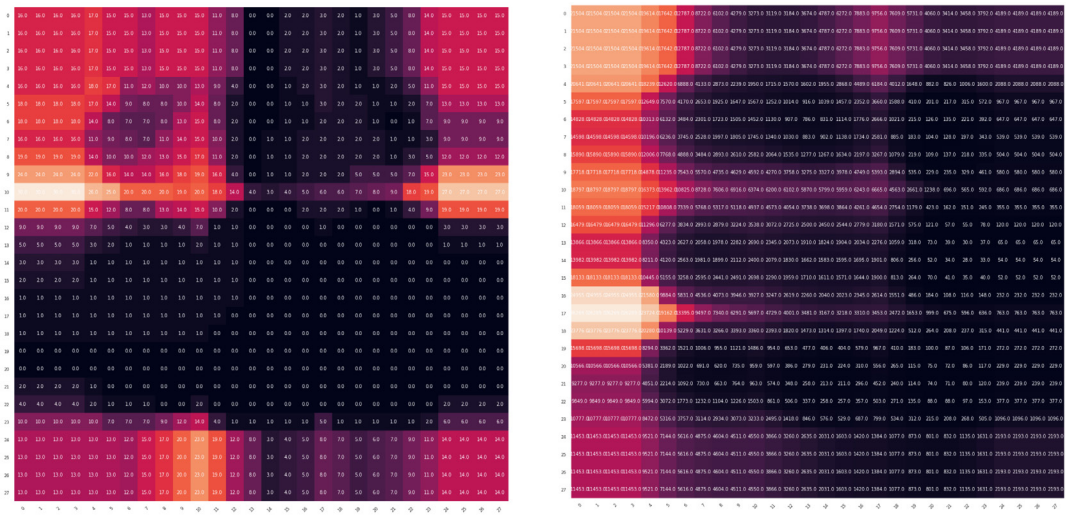
Since the GASF images are difficult to interpret in terms of Grad-CAM, a new approach for creating and combining focused Grad-CAM images has been developed, in this paper. Thus, it is aimed to objectify the dominant relationships in decision of CNN. First of all, to distinguish between feature relationships that have a greater effect on classification, thresholding is applied to images. The value of each pixel, where the image density is less than a fixed value, is changed to black. Thus, binary masks with focused Grad-CAM results are obtained by filtering the relationships above a certain threshold value ∅ in the heat map. Focused Grad-CAM images, is calculated as given in Eq. 20.

$$F^c(i,j) = \begin{cases} 1, & L^c(i,j) \geq \theta \\ 0, & other \end{cases} \tag{20}$$

Finally, focused Grad-CAM images of each class in the test set were combined to identify and visualize the relationships deemed most important. In a test set with n focused Grad-CAM images of *mxm* size, $T^c$ which is the highlighting frequency of key pixels, is calculated as given in Eq. 21.

$$T^c(i,j) = \sum_{k=1}^{n} \sum_{i=1}^{m} \sum_{j=1}^{m} k_{i,j}^c \tag{21}$$

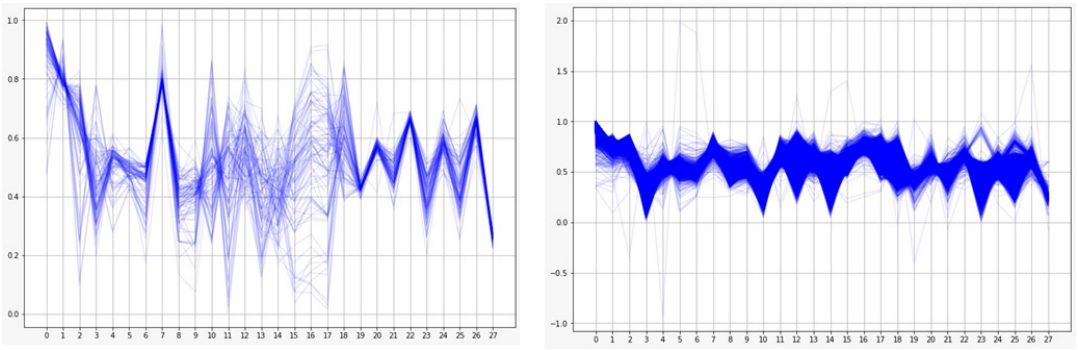
The images produced to reveal fraudulent characteristics and relationships are given in Figure 5. The red or blue areas in the GASF show the areas where the relationship between the two intersecting feature is the strongest while the green and orange areas show moderate relationships. Bright areas in the Grad-CAM output point to the regions that the model focuses on during classification, that is, the most distinctive relationships. The transition from blue to yellow visualizes the transition from the least effective pixels to the most effective pixels. Lastly, focused Grad-CAM is presented as a binary mask that summarizes only the most dominant areas of the Grad-CAM image.



*(a)* *(b)*

*Fig. 6. Regions where the model focuses for fraudulent (a) and legitimate (b) classes in test data set*

All focused Grad-CAM results are combined for each class in the test set to visualize the most important relationships used by the model. The focusing frequency of the model on the pixels is given in Figure 6. The model focuses on the relationship of the first six features with the tenth feature and the last four features with the tenth feature in the decision process for the fraudulent class. In addition, the model makes a decision for the legitimate class according to the relationships between the first five features and the seventeenth feature. Looking at the actual signals of the images given in Figure 7, there is no distinctive pattern of feature relationships just mentioned. This is an indication that the hidden patterns in the data have been modeled.



*(a)* *(b)*

*Fig. 7. Original signals of fraudulent (a) and legitimate (b) classes in test data set*

# 4. Conclusion

In this study, a different perspective has been developed for the credit card fraud detection problem. A novel approach named FDIC is proposed. Credit card transactions are considered as time series and converted into images by means of GASF, which is a polar coordinate transformation technique. Then, these images are classified with CNN as fraudulent or legitimate. While evaluating the success of FDIC, $F_1$-score is taken as the base metric because credit card fraud detection data is highly class imbalanced and $F_1$-score provides more realistic results by taking both false positives and false negatives into account. When the obtained results are compared with the related studies, FDIC has the best $F_1$-score, which is 85.49%. This indicates that FDIC is very good at detecting actual positive, which is a fraudulent transaction and associated with high cost. Finally, the results of FDIC were evaluated in terms of explainability. Since the GASF images are difficult to interpret, a new approach has been developed to define the importance of bilateral relations. The regions where the model focuses on both fraudulent and legitimate classes during the classification are visualized on the heat map.

The future proposal might be that the present work can be carried out in the any field which can be expressed as time series. Large time series can be costly in terms of hardware and time at the stages of converting data into images and classifying these images. However, FDIC make it possible to model the positive or negative effect of bilateral relationships between features on the label, unlike classical methods that model the effect of the feature on the label. In addition, FDIC reveals the improvement achieved by simply moving the problem to a different space without the need for manual intervention.

As we move towards more augmented analytics in the age of big data where fraud and legitimate activities are increasing day by day, the explanation of the developed models and insights has become critical in terms of trust, legal compliance, and brand reputation management. For this reason, explainable artificial intelligence, which highlights the strengths and weaknesses of the developed models, predicts their possible behavior and identifies possible biases, which will help build more trust between organizations and their customers and stakeholders.

# 5. References


Abakarim, Y., Lahby, M., and Attioui, A. (2018). An Efficient Real Time Model For Credit Card Fraud Detection Based On Deep Learning. ACM International Conference on Intelligent Systems: Theories and Applications, Rabat, Morocco. 1-7.

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160.

Adewumi, A. O., and Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. International Journal of System Assurance Engineering and Management, 8(2), 937-953.

Al-Shabi, M. A. (2019). Credit Card Fraud Detection Using Autoencoder Model in Unbalanced Datasets. Journal of Advances in Mathematics and Computer Science, 33(5), 1-16.

Arif, M., and Dar, A. R. (2015). Survey on fraud detection techniques using data mining. International Journal of u-and e-Service, Science and Technology, 8(3), 165-170.

Bagnall, A., Lines, J., Bostrom, A., Large, J., and Keogh, E. (2017). The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery, 31(3), 606-660.

Behdad, M., Barone, L., Bennamoun, M., and French, T. (2012). Nature-inspired techniques in the context of fraud detection. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 42(6), 1273-1290.

Clevert, D. A., Unterthiner, T., and Hochreiter, S. (2015). Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Estebsari, A., and Rajabi, R. (2020). Single Residential Load Forecasting Using Deep Learning and Image Encoding Techniques. Electronics, 9(68), 1-17.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. A. (2019). Deep learning for time series classification: a review. Data Mining and Knowledge Discovery, 33(4), 917-963.

Fu, T. C. (2011). A review on time series data mining. Engineering Applications of Artificial Intelligence, 24(1), 164-181.

Gold, S. (2014). The evolution of payment card fraud. Computer Fraud & Security, 2014(3), 12-17.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T. (2018). Recent advances in convolutional neural networks. Pattern Recognition, 77, 354-377.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

Guo, H., Liu, H., Wu, C., Zhi, W., Xiao, Y., and She, W. (2016). Logistic discrimination based on G-mean and F-measure for imbalanced problem. Journal of Intelligent and Fuzzy Systems, 31(3), 1155-1166.

Gupta, D., and Rani, R. (2019). A study of big data evolution and research challenges. Journal of Information Science, 45(3), 322-340.

Hong, Y. Y., Martinez, J. J. F., and Fajardo, A. C. (2020). Day-Ahead Solar Irradiation Forecasting Utilizing Gramian Angular Field and Convolutional Long Short-Term Memory. IEEE Access, 8, 18741-18753.

Hsueh, Y., Ittangihala, V. R., Wu, W. B., Chang, H. C., and Kuo, C. C. (2019). Condition monitor system for rotation machine by CNN with recurrence plot. Energies, 12(3221), 1-13.

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Liu, X., Cai, H., Zhong, R., Sun, W., and Chen, J. (2020). Learning Traffic as Images for Incident Detection Using Convolutional Neural Networks. IEEE Access, 8, 7916-7924.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. ISPRS journal of photogrammetry and remote sensing, 152, 166-177.

Maimon, O., and Rokach, L. (2005). Data mining and knowledge discovery handbook, Boston: Springer, 1069-1103.

Mohammed, R. A., Wong, K. W., Shiratuddin, M. F., and Wang, X. (2018). Scalable machine learning techniques for highly imbalanced credit card fraud detection: a comparative study. Pacific Rim International Conference on Artificial Intelligence, Nanjing, China, 237-246.

Nilson Report 1164. Retrieved on December 1, 2020, from: https://nilsonreport.com/publication_newsletter_archive_issue.php?issue=1164.

Qin, Z., Zhang, Y., Meng, S., Qin, Z., and Choo, K. K. R. (2020). Imaging and fusing time series for wearable sensor-based human activity recognition. Information Fusion, 53, 80-87.

Rusiecki, A. (2019). Trimmed categorical cross-entropy for deep learning with label noise. Electronics Letters, 55(6), 319-320.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, 618-626.

Seyedhossein, L., and Hashemi, M. R. (2010). A Timelier Credit Card Fraud Detection by Mining Transaction Time Series. International Journal of Information and Communication Technology, 2(3), 21-28.

Université Libre de Bruxelles (ULB) Machine Learning Group, Credit Card Fraud Detection Dataset, Retrieved on December 1, 2020, from: https://www.kaggle.com/mlg-ulb/creditcardfraud.

Wang, Z., and Oates, T. (2015a). Imaging time-series to improve classification and imputation. ACM International Conference on Artificial Intelligence, Buenos Aires, Argentina, 3939-3945.

Wang, Z., and Oates, T. (2015b). Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, Austin, United States, 1-7.

Wilson, S. J. (2017). Data representation for time series data mining: time domain approaches. Wiley Interdisciplinary Reviews: Computational Statistics, 9, 1-6.

Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., and Jiang, C. (2018). Random forest for credit card fraud detection. IEEE International Conference on Networking, Sensing and Control, Zhuhai, China, 1-6.

Zhang, R., Zheng, F., and Min, W. (2018). Sequential Behavioral Data Processing Using Deep Learning and the Markov Transition Field in Online Fraud Detection. ACM SIGKDD Conference on Knowledge Discovery and Data Mining Data Science in Fintech Workshop, 1-5.