International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Financial Fraud Detection of Digital Transaction Using Artificial Intelligence & Machine Learning

Shivam Sundaram[1], Nanhay Singh[2]

[1]Department of Computer Science and Engineering, Netaji Subhas University of Technology, New Delhi, India
shivam98.iitd[at]gmail.com

[2]Professor, Department of Computer Science and Engineering, Netaji Subhas University of Technology, New Delhi, India
nsingh1973[at]gmail.com

Abstract: ***The rapid increase in the field of digitalization in the global platform, we Homo sapiens have been highly dependent and rely on such platforms to get our job done in day today work style. With that set we can't also deny the fact some loop holes are also there which can cause a blunder in the ecosystem of financial platform. Yes your guess is right we are talking about the fraud which occurs with the great gift of ease in work through digitalization. Frauds are volatile and dynamic in nature as well as they depicts no sign of pattern therefore it's very difficult to identify. Counterfeiters uses the recent modern tools and advance technology to their advantage. They somehow finds the gaps and succeed in breaching the security checkpoints causing loss of millions of dollars. The major concern is that financial frauds are not limited to cards, Unified Payments Interface and other transactions it deep down goes to websites phishing and many more. Analysis and detection of such anomalies using advance techniques of machine learning and artificial intelligence is possible. In order to avoid such occurrences of anomalies or frauds, this research intends to benchmark several artificial intelligence, deep learning, and machine learning algorithm concepts. The dataset used would be primary, secondary and mostly random datasets available across the various sources encountered during period of research.***

Keywords: RFbFFD (Random Forest based Financial Fraud Detection), Cards, frauds, Digitalization

1. Introduction

This In the era of digitalization and globalization there has been tremendous growth in the internet. Which has a huge impact on the diversity and modern lifestyle influencing the way of payments, ease of not carrying a bulk of cash and it have liquidated the flow of financial transactions. The impact to the above stated is that there is an enormous increase in the rate of cases seen day by day in the financial fraud with respect to cyber-attacks and Several con artists have discovered methods to take advantage of consumers and ransack their financial data in order to utilize it for unauthorized transactions ever since bank cards and contactless transactions were first introduced. This results in a significant number of fraudulent purchases each and every day. [1]. But there are numerous ways to protect and bind these activities to prevention including data encryption and tokenization using machine learning techniques and artificial intelligence algorithm which will help in identifying, decision making. And blocking such activities actively and passively. Although some methods working alone would not be enough to put into container but combination of such can be very productive to various extent. Machine learning is sub field of artificial intelligence which has the ability to learn from the past experience and helps in judgments for future instances without much of manual and human intervention. In addition to the harm done to their brand, image, and market share, respondents to PricewaterhouseCoopers (PwC's) worldwide Economic fraud and crime Study 2022 estimated total losses of US$42 billion. The bulk of the fraud was committed on the outside, such as through e-commerce, 31% by hackers, etc. [2].

Therefore we can see that it's very pivotal to implement a proper system which can help in detecting the financial losses and preventing the user from such damage through protecting. The most difficult part in the implementation of this model through machine learning techniques is that as we know financial transactions are highly confidential getting that dataset is nearly impossible from the financial institutions. With that respect we are left out with one and only options that are freely available across the internet which have anonymized attributes. One option is that we can create multiple dataset after studying various dataset which are exactly similar to the financial institution where we can get some promising attribute but this would be not enough to train the model. Furthermore this would be not quite easy task as there is a constant change in the techniques and patterns of fraudulent activities by the fraudsters. Addition to this existing machine learning models are of low accuracy level and are not able to solve and take right judgment on the highly skewed nature of fraud datasets. Therefore there is a need of machine learning with deep learning model to optimize such situation with high accuracy.

a) Challenges

- Extremely high false positives can be encountered when the detected data is copy of copy in other words we can say when a fraudster is also having false identity that could cause problem in detecting exact source of scam.
- Missing Omni channel coverage cause various problem like exact data and multiple source of transactions are not tracked out at one system software to aggregate into a single source is a hectic process.
- Delayed alert mechanism is another set of issue which we can see in most of the cases and research paper till date. Once a fraud transaction is made then that data is processed into bulk need time to detect and alert the institution or system.
- Extremely high false positive and missing insight from unstructured data is caused as with the intelligent system fraudster are also getting smarter and ways to trick the system and not get tracked so they find different methods to breach the system in the unstructured way not letting to easily detect the pattern which causes the high false

positive data and not a particular attribute which can be used to track the data.
- Need for business user driven and lack of insight from analytics – This is most challenging task to integrate all the methodologies into a single platform or channel for a best solution which is business driven and can help us in analysis of the fraud being done in the real time and produce a best track into a single platform.

b) Motivation

There are multiple factors which let me to initiate the research on the field of fraud detection some of them are illustrated below:

- In the current scenarios n numbers of transaction are being done worldwide among them some transactions are hoax which causes middle class people there hard earned money loose in just a few steps causing them a huge damage into their financial status and survival of their livelihood.
- Finally we don't have any sort of stable and good solution towards the situation faced though we have good technology these days still a proper measure to tackle the counterfeiter which can help in prevention is all platform is not present.
- All projects and research are there to help but somehow at certain stage dynamic solution to the problem and coverage of all single point is not feasible. But exploration to achieve specific solution to surpass the difficulty can be done at minimization of loss.
- Attaining knowledge to a specific field and finding decent solution to help the targeted society is most powerful inspiration which drives attention to work on this field.

2. Related Work

A lot of work has been undertaken in various directions regarding fraud detection especially in the credit card section and e-commerce transaction and these works are available on the web and a lot more is still going on. Let's have the comparative analysis of all the findings and research done based on journal's, research paper published, articles and web content in tabular form so that each finding can be truly compared and layman to understand.

The findings are done on the random data and the accuracy is done on the relative basis of the techniques utilized such as K-means, SVM and ANN. Basically it distinguish between fraudulent and legitimate transactions, they essentially did a distinction of the successful monitored expert system approaches [3]. Alert system was introduced in this paper that's a real time process and causes the prevention measures as well. Through comparison between all different techniques it was found that DCNN bypass all other techniques like SVM, LR, RF and RNN with more speed and accuracy as per verdicts by the author of paper [4]. While building a model to identify credit card fraud, the authors of this paper suggest a sequence based classifier on LSTM networks to track the consumer behavior of individual cards. [5].

The paper basically deals with performance comparison with respect to the different kind of dataset which can be useful for industry point of view. Based on studies by the author, it has been determined that support vector machines are the best way for detecting fraud with large datasets, maybe when paired with CNN to get more credible results. Combining SVM, Random Forest, and KNN techniques can produce effective improvements for smaller datasets. Despite the fact that supervised learning techniques like CNN, KNN, and Random Forest are appealing and have high outcomes, they do not function well in dynamic contexts. Auto encoders are an excellent solution in such case as they're only trained on ordinary (i.e., non-fraudulent) traffic. When transactions deviate from the expected patterns, they are flagged as fraudulent. Even when auto encoder training is although originally rather expensive, it can be helpful for data labelling sets. A sufficient amount of labelled data can be utilized to retrain or create further supervised models [1]. The method this study suggests makes use of the most recent machine learning techniques to identify unusual behaviors, often known as outliers. The following figure 1 can be used to depict the fundamental rough architectural diagram.

Figure 1: Fundamental Architectural Diagram

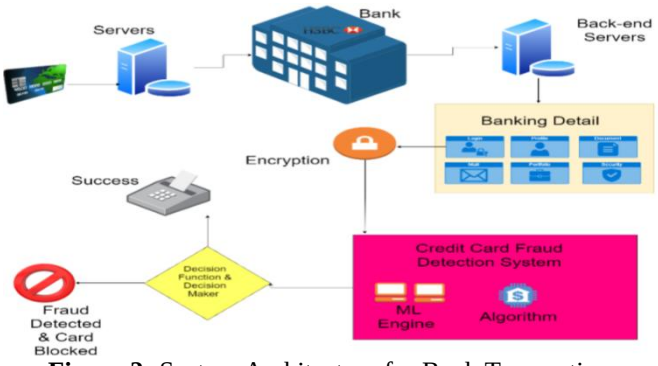


Figure 2: System Architecture for Bank Transaction

When we look precisely on the larger picture the architecture looks like above Figure 2. While the precision remains 28% only to improve accuracy rate this is again a big sort of problem for larger set of data in this model [7] (figure 1& 2 reference [7]). The major change in the paper was it can handle all concurrent data where swarm intelligence is utilized to optimize a subset of pertinent feature. The Synthetic Minority Oversampling Method (SMOTE) was developed to address the issue of unbalanced data, while the UMAP approach was used to decrease dataset dimensionality. The long-term dependencies inside the transaction sequence were removed using LSTM.

This model is capable of recognize the useful pattern with respect to customer behavior which helps in detection of fraud from normal transactions [8]. The basic understanding on the comparative analysis of different algorithm with combination to four different dataset proved that the best

possible method after applying on all dataset Gradient Boosted Trees (GBT) surpassed all other algorithm with average accuracy of 96% [9].

a) Problem Definition

There are multiple instances which we can observe while we go through different research paper. Each research analysis defines and states the different issue observed while conducting the solution to the problem statement. Major issue which was observed in majority of the research and the major challenging part is that there is no proper or exact dataset in the internet market and obtaining the dataset from the financial institution is nearly not possible due to privacy issue. Some other issue which was seen while findings are that no proper attributes can be selected for a dataset as this occurrences is not static in nature with various fraud the pattern are different, dynamic and attribute selection can be different, so we need a inline solution for this were all the major issues of selection can be satisfied. Selection of exact algorithm that can handle large dataset and can help in permutation and combination of such past history to obtain the new result and can also be helpful in prediction of future instance of fraud would be helpful. The platform dependencies which was the other issue like for e-commerce transaction, banking transaction and many more should be solved at single platform.

b) Research Objectives

The research work proposes the different types of permutation and combination of algorithmic approach to observe the better instances to solve the problem definition and to provide the better solution for financial fraud detection and prevention. The following objective needs to be persuade in order to satisfy the problem statement solution:

- Create the database or folder for different set of financial data.
- Feature selection and extraction of data on the basis of best possible attribute for the give case.
- Create a model to train the dataset and give effective response to generate the accurate result, though decision making on the basis of evaluation metrics dataset.
- Redefining the model for predictive analysis and also comparative analysis of the each model defined.
- Finding conclusion on the basis of observations.

3. Research Methodology

I'm willing to create a best possible solution with high accuracy and can be applied to all possible dataset and can also help in prevention as well as minimization of such fraudulent transactions. The various techniques with all possible combination would be applied so that best possible case can be encountered without over fitting and under fitting of data. The flow diagram for the process are as follows:

- Gathering of dataset- (Minimum 2-4)
- Feature selection and extraction – dimension reduction
- Algorithmic approach utilized would be Recursive Feature Elimination (RFE), PSO (particle swarm optimization) and Gradient Boosting Algorithm and AdaBoosting Algorithm.
- Feedback Mechanism to update and make system smarter for future transaction.

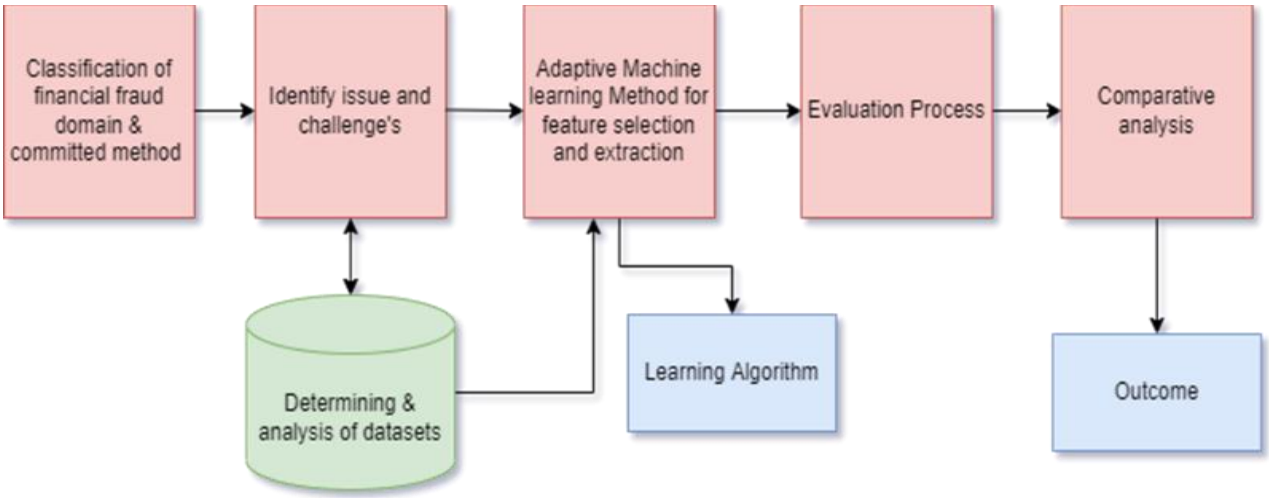


Figure 3: Representation of flow diagram as per research methodology.

Further modification in flow diagram would be represented in the future journey of research process. All possible algorithms will be discussed with implementation and result in other section of related work. Alert system design would be basic as discussed in point four as feedback mechanism. This process will be initiated from learning algorithm block.

4. Implementation

The implementation of the work majorly focus on the financial dataset. We all know that gathering financial data is how much crucial and it's always the primarily focus of the financial institution to keep their customer information confidential with utmost priority. Increase in the transaction comprises with high volume of data bundles and handling such data is also very complex process their might be similar multiple transactions and recognitions of such patterned transaction finding the scope of fraudulent activity is nearly impossible with the advancement of technology and new high tech equipment device. So training a model on single algorithm for a dataset would be not practically fruitful, we need a smart hybrid model which can handle irregular activity efficiently.

a) Dataset amd Features

- Dataset is of 150.6 MB in size.

- Total transaction is of 48 hrs. (2 days)
- Number of transaction is 284807 in numbers and two lacs eighty four thousand eight hundred and seven.
- Total number of features or attributes is 31, including labeled data features.
- Class features 0 defines genuine transaction and 1 indicates fraud transaction.
- Among 31 features 3 attributes have be taken into major consideration for evaluation of the process where it can justify the transaction is genuine or fraud, which are Amount, Class and Time.
- Dataset is in comma separated format (CSV).
- Features which are highly confidential like card details, transaction details, location, type of truncation, gender, citizenship, profession, salary type, employment etc. are replaced by numerical values by the provider or data owner.

Visualization of dataset .csv file and graphs can be visualized in figure 4 and 5:

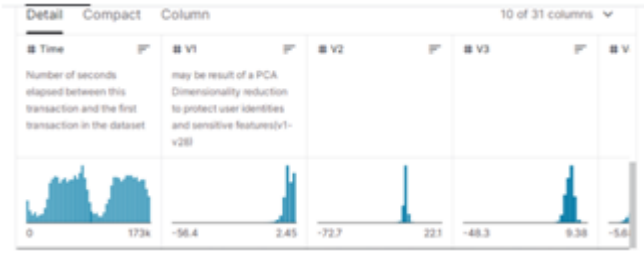


Figure 4: Representation of data on the basis of Graph Plotting

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010

5 rows × 31 columns

Figure 5: Dataset in tabular CSV format.

The major focus is to test different kind of financial dataset, at least 4-5 different dataset which can altogether can set up the benchmark to get the good findings.

b) Methodology Implication

The code has been implemented using python programming language and various libraries has been imported for utilization like numpy, pandas, sklearn, pycaret, matpoltlib. User interface and IDE used is Jupiter notebook and Google Colab.Basic start on code goes with importing of libraries and importing csv file to read suing pandas libraries. Command used to load in pandas dataframe is Name = pd.read_csv('Path of file').Printed first 5 rows of dataset to check its loading perfectly using command Name.head(). Similarly last 5 rows can be printed using Name.tail(). Next process is to check data information whether data does have proper formatting and no missing values are present.

5. Results

The processing result shows no missing value or null value in the dataset which is shown below in Figure 6.

```
[ ]  RangeIndex: 284807 entries, 0 to 284806
     Data columns (total 31 columns):
      #   Column  Non-Null Count   Dtype
     ---  ------  --------------   -----
      0   Time    284807 non-null  float64
      1   V1      284807 non-null  float64
      2   V2      284807 non-null  float64
      3   V3      284807 non-null  float64
      4   V4      284807 non-null  float64
      5   V5      284807 non-null  float64
      6   V6      284807 non-null  float64
      7   V7      284807 non-null  float64
      8   V8      284807 non-null  float64
      9   V9      284807 non-null  float64
      10  V10     284807 non-null  float64
      11  V11     284807 non-null  float64
      12  V12     284807 non-null  float64
      13  V13     284807 non-null  float64
      14  V14     284807 non-null  float64
      15  V15     284807 non-null  float64
      16  V16     284807 non-null  float64
      17  V17     284807 non-null  float64
      18  V18     284807 non-null  float64
      19  V19     284807 non-null  float64
      20  V20     284807 non-null  float64
      21  V21     284807 non-null  float64
      22  V22     284807 non-null  float64
      23  V23     284807 non-null  float64
      24  V24     284807 non-null  float64
      25  V25     284807 non-null  float64
      26  V26     284807 non-null  float64
      27  V27     284807 non-null  float64
      28  V28     284807 non-null  float64
      29  Amount  284807 non-null  float64
      30  Class   284807 non-null  int64
     dtypes: float64(30), int64(1)
     memory usage: 67.4 MB
```

Figure 6: Shows no null value present in any features

Next process is distribution of legit transactions & fraudulent transactions. This will help in acknowledging either data is balanced or not.

```
[ ]  1 # distribution of legit transactions & fraudulent transactions
     2 credit_card_data['Class'].value_counts()

0    284315
1       492
Name: Class, dtype: int64
```

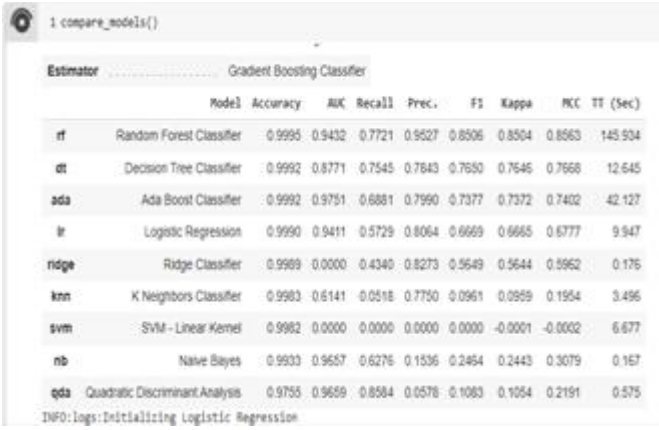
This Dataset is highly unblanced

```
1 # statistical measures of the data
2 legit.Amount.describe()

count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

```
1 fraud.Amount.describe()

count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

Figure 7: Legit and fraud transaction summary

TheFigure 7 represent the description of authorized and hoax transaction in the data file using the written command in the picture and also we got to know through result that data is highly imbalanced as the ratio of legit vs. fraud transaction have huge difference in numbers. Splitting the data into features and target is required so there is need of co-ordinates. Once all these process have taken place then we need to train and test the data.



```
1 compare_models()
```

Estimator ..... Gradient Boosting Classifier

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.9995	0.9432	0.7721	0.9527	0.8506	0.8504	0.8563	145.934
dt	Decision Tree Classifier	0.9992	0.8771	0.7545	0.7643	0.7650	0.7646	0.7668	12.645
ada	Ada Boost Classifier	0.9992	0.9751	0.6881	0.7990	0.7377	0.7372	0.7402	42.127
lr	Logistic Regression	0.9990	0.9411	0.5729	0.8064	0.6669	0.6665	0.6777	9.947
ridge	Ridge Classifier	0.9989	0.0000	0.4340	0.8273	0.5649	0.5644	0.5962	0.176
knn	K Neighbors Classifier	0.9983	0.6141	0.0518	0.7750	0.0961	0.0959	0.1954	3.496
svm	SVM - Linear Kernel	0.9982	0.0000	0.0000	0.0000	0.0000	-0.0001	-0.0002	6.677
nb	Naive Bayes	0.9933	0.9657	0.6276	0.1536	0.2454	0.2443	0.3079	0.167
qda	Quadratic Discriminant Analysis	0.9755	0.9659	0.8584	0.0576	0.1083	0.1054	0.2191	0.575

```
INFO:logs:Initializing Logistic Regression
```

Figure 8: Results of Confusion Matrix

According to Figure 8, Random Forest Classifier shows the greater accuracy in terms of large datasets. The kappa value is near to 1, so it states that there is a perfect agreement between the raters. The RFbFFD model outperforms all the other classification methods in terms of evaluation metrics.

6. Conclusion and Future Scope

The scope of this research would be very much impactful as multiple sources are present and no such work has been done before for prevention of system the major work has been only implemented in the domain of detection of fraud cases using machine learning approach. It will provide easier in understanding the traction process as well as it will help in filling up the gap areas such as dynamic fraud detection analysis for various platforms not limiting to only credit card fraud analysis but also to various other platform such as e-commerce and other financial sector. Major focus would be to get the deterministic result for different dataset that can overcome all other previous research scenarios with maximization of Precision and Accuracy.

References

[1] Pradheepan Raghavan and Neamat El Gayar, Fraud Detection using Machine Learning and Deep Learning. Page. 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) December 11-12, 2019, publication at: https://www.researchgate.net/publication/339411416

[2] https://www.pwc.com/gx/en/services/forensics/economic-crime-survey.html. Data based on the pwc website for survey 2022.

[3] Asha RB and Suresh Kumar KR. Credit card fraud detection using artificial neural network. Global Transitions Proceedings 2 (2021) 35–41, www.elsevier.com/locate/gltp

[4] Joy Iong-Zong Chen and Kong-Long Lai, Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert publications: Journal of Artificial Intelligence and Capsule Networks (2021) Vol.03/ No.02 Pages: 101-112, http://irojournals.com/aicn/ DOI: https://doi.org/10.36548/jaicn.2021.2.003

[5] Ibtissam Benchaji, Samira Douzi, and Bouabid El Ouahidi, Credit Card Fraud Detection Model Based on LSTM Recurrent Neural Networks, Journal of Advances in Information Technology Vol. 12, No. 2, May 2021

[6] Ameer Saleh Hussein, Rihab Salah Khairy, Shaima Miqdad Mohamed Najeeb and Haider Th. Salim ALRikabi, International Journal of Interactive Mobile Technology(iJIM)- eISSN: 1865-7923 –Vol. 15, No. 05, 2021

[7] S P Maniraj, Aditya Saini, Swarna Deep Sarkar and Shadab Ahmed, Credit Card Fraud Detection using Machine Learning and Data Science, International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV8IS090031 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by: www.ijert.org Vol. 8 Issue 09, September-2019.

[8] Ibtissam Benchaji , Samira Douzi, Bouabid El Ouahidi and Jaafar Jaafari, Enhanced credit card fraud detection based on attention mechanism and LSTM deep model, Benchaji et al. Journal of Big Data (2021) 8:151 https://doi.org/10.1186/s40537-021-00541-8.

[9] Manjeevan Seera, Chee Peng Lim, Ajay Kumar, Lalitha Dhamotharan, Kim Hua Tan, An intelligent payment card fraud detection system, Accepted: 3 June 2021© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

[10] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[11] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.