Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin

Conference Paper · June 2020

DOI: 10.1145/3409073.3409078

3 authors:

Ismail Alarab
Bournemouth University
10 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Simant Prakoonwit
Bournemouth University
48 PUBLICATIONS **265** CITATIONS
SEE PROFILE

Mohamed Ikbal Nacer
Bournemouth University
8 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin

Ismail Alarab
Bournemouth University
United Kingdom
ialarab@bournemouth.ac.uk

Simant Prakoonwit
Bournemouth University
United Kingdom
sprakoonwit@bournemouth.ac.uk

Mohamed Ikbal Nacer
Bournemouth University
United Kingdom
mnacer@bournemouth.ac.uk

ABSTRACT

With the advance of Bitcoin technology, money laundering has been incentivised as a den of Bitcoin blockchain, in which the user's identity is hidden behind a pseudonym known as address. Although this trait permits concealing in the plain sight, the public ledger of Bitcoin blockchain provides more power for investigators and allows collective intelligence for anti-money laundering and forensic analysis. This fascinating paradox arises in the strength of Bitcoin technology. Machine learning techniques have attained promising results in forensic analysis, in order to spot suspicious behaviour in Bitcoin blockchain. This paper presents a comparative analysis of the performance of classical supervised learning methods using a recently published data set derived from Bitcoin blockchain, to predict licit and illicit transactions in the network. Besides, an ensemble learning method is utilised using a combination of the given supervised learning models, which outperforms the given classical methods. This experiment is performed using a newly published data set derived from Bitcoin blockchain. Our main contribution points out that using ensemble learning approach outperforms the performance of the classical learning models used in the original paper, using Elliptic data set, a time series of Bitcoin transaction graph with node transactions and directed payments flow edges. Using the same data set, we show that we are able to predict licit/illicit transactions with an accuracy of 98.13% and F_1 score equals to 83.36% using the proposed method. We discuss the variety of supervised learning methods, and their capabilities of assisting forensic analysis, and propose future work directions.

CCS Concepts

• **Computing methodologies→Ensemble Methods; Supervised learning by classification** • **Security and privacy→Database activity monitoring**

Keywords

Ensemble Method, Supervised Learning, Anomaly Detection, Financial Forensics, Anti-Money Laundering

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

1. INTRODUCTION

Money laundering has received meticulous attention with the emergence of cryptocurrencies. Criminals have perceived the Bitcoin network as an advanced process to promote their competences. Bitcoin blockchain has been identified as a peer-to-peer decentralized bank for the Bitcoin cryptocurrency [1]. In contrast to normal banks, the transactions in Bitcoin are interpersonal, which are digitally signed and verified in a public ledger without any intermediaries. In Bitcoin blockchain, transactions are processed between addresses which are derived from the public and private keys of the users' wallets. Primarily, Bitcoin addresses are not associated with the individual identity, in which the user is hidden behind a pseudonym. On the other hand, the historical information of any transaction or Bitcoin address is public, and any transaction is linked to the previous ones [2]. For instance, the output of clean money which is originated from the money laundering process can be traced due to the nexus Bitcoin blockchain.

The advent of Bitcoin blockchain has provided a mysterious intriguing technology, between high anonymity (commonly known as pseudo-anonymity) and public availability of Bitcoin transactions. Due to pseudo-anonymity and untraceability of Bitcoin, it is currently used by criminals in illegal activities such as money laundering and mixing services [3]. For this reason, financial regulators, law enforcement, intelligence companies who use Bitcoin blockchain have become aware of technical developments in societal adoption of the cryptocurrency Bitcoin [4]. Banks are subjected to Know-Your-Customer (KYC) principle, which is a mandatory requirement of the individuals to validate the identity of account holders [5]. But in the public Bitcoin ledger, the addresses are pseudonyms unless they are associated with the identity information [6].

Cryptocurrency intelligence companies have exploited the public ledger of cryptocurrency such as Bitcoin blockchain to provide Anti-Money Laundering (AML) solutions according to the cryptocurrency domain. Elliptic company is a cryptocurrency intelligence company that has provided the publicly available Elliptic data. This data set is a graph network of Bitcoin transactions, which is considered as one of the largest labelled data set available in any cryptocurrency [7]. This data set is highly imbalanced data and only 2% of the transactions are illicit that justify illicit services, while 21% are licit that describe normal transactions, where the rest are of unknown labels. In case of imbalanced data set, it is desired to improve the recall, while preserving the precision of the model. However, the latter two terms cannot be optimized simultaneously; an increase in one of these two terms may lead to a decrease in the other term, since increasing true positives might increase the false positives at the same time, hence reducing the precision.

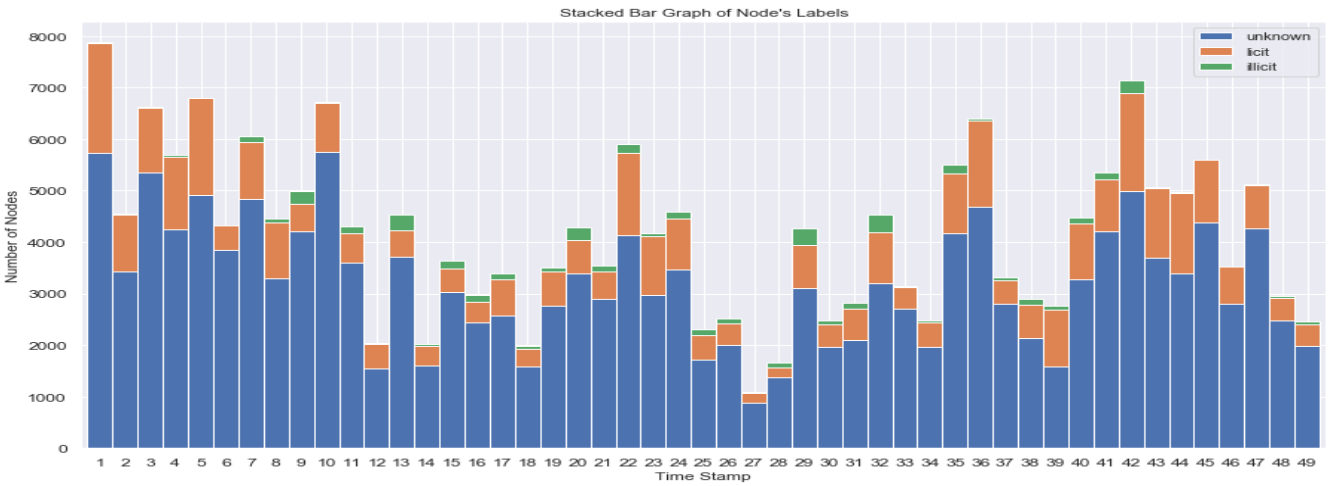

Figure 1: Distribution of the number of nodes according to the timestamps.

The combination of precision and recall is so-called F_1 score, which can indicate the trade-off between the two former metrics, and reflect the goodness of classification in the model [8].

For this purpose, we aim to provide a comparative analysis on Elliptic data set using a variety of supervised learning techniques, in which the goodness of classification of our models performs well in comparison to the previous research in [7]. Moreover, we perform ensemble learning, a combination of machine learning predictors [9] that prevails other classical learning methods at predicting licit/illicit transactions. In our experiment, ensemble learning can be defined as a classification method based on average probability ensemble [10], derived from the collection of best performing supervised learning methods used in our experiment.

This paper can be roughly divided into the following sections. In Section 2, we give an overview of the related work. Section 3 states existing methods used in our experiment, while Section 4 provides the experiments and the results are interpreted in Section 5. The conclusion with future work is provided in Section 6.

2. OVERVIEW OF RELATED WORK

2.1 Elliptic data set

Regarding the used data, Elliptic data belongs to real Bitcoin transactions, which form a directed graph network consisted of nodes representing Bitcoin transactions and edges representing payments flow from the source to the destination. This data set is classified into two categories licit and illicit transactions. The licit category belongs to Bitcoin mining, exchanges, wallet provider, licit services, etc. However, the illicit category is associated with illegal transactions such as thefts, scams, malware, ransomware, etc. This data set consists of 49 timestamps uniformly spaced with an interval of two weeks, and each time-step represents a distinct collection of transactions to form a single connected digraph that has appeared within less than three hours in the blockchain [7].

There are 203,769 nodes transactions and 234,355 directed edges of the payments flow, where 2% (4,545) of the nodes transactions are labelled as illicit transactions, and 21% (42,019) of nodes transactions are licit. However, the remaining transactions are enriched with nodes features but with unknown labelling. The nodes of the graph network are formed of 166 features which are constructed of only publicly available information [7]. The first 94 features of the nodes belong to the local information of the Bitcoin transactions such as time-step, transaction fees, number of input/output ...etc., where the remaining 72 features represent the aggregated information obtained from one-hop backward/forward aggregation of graph nodes, which are associated with the structural information of the graph network as forward from the center node, giving the maximum, minimum, standard deviation and correlation coefficients of the neighbour transactions for the same information data (number of inputs/outputs, transaction fee, etc.) [7]. The distribution of the number of nodes labelling according to different time-step is shown in Figure 1 thanks to [11].

2.2 Previous Research

The prominence of forensic analysis in Bitcoin blockchain has widely arisen with the advance of blockchain technology, in which the criminals tried to exploit Bitcoin for illicit services. Due to the complexity of Bitcoin data, different methods have been widely investigated to explore different activities done on the ledger. In [3], "BlockchainVis" has been introduced as a visual analytics tool, to filter out non-interesting information, and to visually analyse specific characteristics in the Bitcoin blockchain. Dealing with a numerous number of nodes and edges in big data is a tricky problem, whilst the target is only to spot illegal services in the network for further analysis. Consequently, a straightforward visualisation is not an appropriate way to visualise the Bitcoin graph network to only analyse the suspicious behaviour. The emerging of intelligent methods such as machine learning techniques can mitigate this weakness based on the historical data faced by normal and illicit activities, where the interesting nodes could be the illicit activities penetrating the Bitcoin blockchain network. The exploitation of machine learning methods in analysing Bitcoin network has successfully revealed promising results. For instance, the complex data of Bitcoin blockchain can be useful for machine learning method to train a model rather than analysing the blockchain data manually. Based on the trained model, the new unseen nodes can be predicted and further analysis could be done by an expert, by visualising the interesting nodes. Another contribution in [12] has performed unsupervised learning using transaction and user graphs of Bitcoin blockchain data. The work in [12] used different clustering methods such as k-means and Gaussian mixture models on a data set derived from Bitcoin blockchain, to detect anomalies or suspicious behaviour without any confirmation if these nodes are conducting illicit activities, in which the model was not very effective [12]. In [4], various supervised learning methods were applied to classify the non-identified clusters in the Bitcoin blockchain network, which have provided acceptable outcomes. The latter research has investigated

several supervised learning techniques to pave the way towards detecting high-risk transactions.

Currently, our experiment is done in light of the previous research in [7], using Elliptic data. This data has been introduced to be publicly available and labelled licit/illicit node transactions collected from Bitcoin blockchain. The original work in [7] has provided the main contribution of this data set in AML use-case, by applying different machine learning methods to predict the licit/illicit transactions based on the historical Bitcoin data set. The features of Elliptic data set, used in [7], were categorized into three different combinations: local features denoted by LF (the first 94 features), all features denoted by AF (166 features), and all features concatenated with node embedding features acquired from Graph Convolutional Network algorithm denoted by AF+NE. However, the original form of the feature matrix consists of local features and aggregated features, in which the importance of aggregated features was addressed besides the local information in [7].

The different combinations of data set features were applied to classical machine learning algorithms such as Logistic Regression, Multi-Layer Perceptron, Random Forest, and Graph Convolutional Network (GCN). These algorithms were tested and compared among all combinations of the features. Eventually, Random Forest has outperformed GCN and other methods. However, GCN has been originally defined for undirected graph [13], and the additional robustness of this method can be interpreted more by having a weighted adjacency matrix instead of un-weighted one. For instance, in a graph Laplacian, if two nodes are linked with a large weight, then the values of the eigenvectors at those locations are likely to be the same. However, the eigenvectors associated with larger eigenvalues oscillate more rapidly, and are more likely to have different values on vertices connected by an edge with high weight [14], which can interpret the importance of the weighted edges. On the other hand, the given graph network might lack to the necessary patterns in its structure, such as the unlabelled transactions in Elliptic data set. This could reduce the performance of GCN algorithm. The contribution in [7] has presented the outperformance of Random Forest using all features AF. In addition, the concatenation of the embedding features that represent the output of GCN layers, with the original data set features, (AF + NE), has enhanced the performance of the models rather than using AF. Original research has pointed out the issue of the given data, where the shutdown of the dark market at 43^{th} time-step occurred, referring to [7]. This event was demonstrated by a rapid decrease in illicit transactions. The sudden closure of dark market has caused all models to perform poorly at this time-step.

3. METHODS

In this section, we will describe the necessary details of the input data used to be fed into the supervised learning methods, then we will discuss the various supervised learning algorithms, as well as ensemble learning that is exploited to build a classifier and perform the prediction on the binary labels (licit/illicit). We will discuss the necessary arguments used to enhance the performance of the model. We will finally conclude with some data limitations.

3.1 Data Preparation

Following the above-mentioned features of the Elliptic data set, the local features excluding the time-step concatenated with the aggregated features are used in our experiment to enhance performance of the various machine learning methods. Thus, the total number of input features is 165 features which describe the dimensional feature space. The train/test set split is performed following the temporal split as the train set belongs to the first 34 timestamps (from 1 till 34), and the test set belongs to the remaining timestamps (from 35 till 49). Furthermore, the data set is highly imbalanced, and the input of the data prepared for supervised learning methods is only considered for the known labels licit/illicit as shown in Table 1.

According to [7], the labelling process of transactions into licit and illicit has been situated using heuristics based reasoning process. For instance, a higher number of input with the reuse of same address can be mapped to the same entity in the Bitcoin blockchain [15], and provide more benefits in terms of transaction costs (fee); this can reduce the anonymity of the user and is more likely to be licit transactions. In contrast, the users following a low number of addresses are more likely to be illicit, and the strength of addresses clustering is reduced [15].

Table 1: Elliptic data set description.

Transactions	Licit	Illicit	Unknown
Train set	26432	3462	106371
Test set	15587	1083	50834
Total	42019	4545	157205

3.2 Benchmark Methods

In this experiment, we have applied the Supervised Machine Learning algorithms which are popular for the analysis of Bitcoin transaction data, referring to [4], as following:

- Random Forest
- Extra Trees
- Gradient Boosting
- Bagging Classifier
- AdaBoost
- k-Nearest Neighbours

Logistic Regression and Support Vector Machine are excluded from the current experiment as both algorithms do not perform well due to the highly imbalanced data set, in which the boundary decision is skewed toward the majority class (licit transactions), as well as the existence of minorities (illicit transactions) in the neighbourhood of the former class. Also, these algorithms are not suitable for the given data set as they revealed a low performance.

Supervised methods in this experiment focus on anomaly detection task. The challenge here is to identify the criminals in a highly imbalanced growing data set. In terms of machine learning, the aim is to achieve a good classification rule by reducing the false positives (licit transactions detected as illicit), without increasing the false negatives (illicit transactions detected as licit).

4. EXPERIMENT

Regarding the given classical supervised learning algorithms, we used scikit learn package [16] of Python Programming Language in all mentioned models, to perform the classification of licit/illicit transactions of Elliptic data set. We fit the train set with a variety of supervised algorithms, while the test set is used to predict the performance of the model. At first, we tested the performance of Random Forest algorithm (with n_estimators=100,

Table 2: Evaluation of supervised learning methods using Elliptic data.

Model	Accuracy	Precision	Recall	F_1 score	False Positives	False Negatives
Ensemble Learning	**98.13%**	**99.11%**	**71.93%**	**83.36%**	**7**	**304**
Random Forest	98.06%	97.38%	72.20%	82.92%	21	301
Extra Trees	98.01%	98.70%	70.36%	82.15%	10	321
Bagging	98.01%	96.41%	72.11%	82.51%	29	302
AdaBoost	97.99%	96.28%	71.83%	82.28%	30	305
Gradient Boosting	97.35%	99.84%	59.37%	74.46%	1	440
k-Nearest Neighbours	95.10%	61.60%	63.99%	62.77%	432	390

Bootstrap=False, min_samples_leaf=2, max_depth=50), Extra Trees (using the same settings as Random Forest), Gradient Boosting (using learning_rate=0.01, min_samples_leaf=2), AdaBoost algorithm and Bagging classifier (both classifiers using Random Forest model as the base estimator). After that, we applied k-Nearest Neighbour (k-NN) algorithm (using scikit learn) after choosing k=8 as the optimal performance in the range $k \in [1, 26]$. Besides the mentioned methods, ensemble learning is also applied using the combination of the three best-performing methods. Ensemble learning is defined as a classification learning method derived from combining a variety of machine learning algorithms, to enhance the performance of the final predictions [9]. Ensemble learning has been widely investigated in previous researches, for its capability for achieving higher accuracy by using predictions from several learning methods to contribute to the final classifications. In our experiment, we used ensemble learning method based on average probability ensemble as named in [10]. In average probability ensemble, the classification is done by using several pre-trained machine learning models, in which the final predictions are derived from averaging the summation of the prediction probabilities obtained from the learning algorithms each. In our experiment, we performed ensemble learning using the combination of the following methods: Random Forest, Extra Trees and Bagging classifiers. Each of these methods provides output predictions as probability values that demonstrate the confidence of the algorithms at labelling the given input vectors. Admittedly, ensemble learning based on average probability ensemble has outperformed all the classical models used from the benchmark methods.

Using Elliptic data set, we fit the mentioned models, after tuning empirically the model hyper-parameters. We then evaluated the results using different machine learning metrics such as accuracy, precision, recall, F_1 score, as well as the number of false positives and false negatives for the sake of clarity as shown in Table 2. Furthermore, we presented receiver operation curve (ROC) to roughly reveal the performance of the used supervised methods, as well as computing the area-under-curve (AUC) of each model as depicted in Figure 2.

5. DISCUSSION

In this research work, we have done a comparative analysis of different supervised learning algorithms to predict licit/illicit transactions using Elliptic data set. The proposed method of ensemble learning has performed the best in comparison to the variety of used supervised learning methods as provided in Table 2. Our results show that ensemble learning is able to perform classification with accuracy 98.13% and F_1 score 83.36% to predict licit/illicit transactions. Our main finding is that ensemble learning

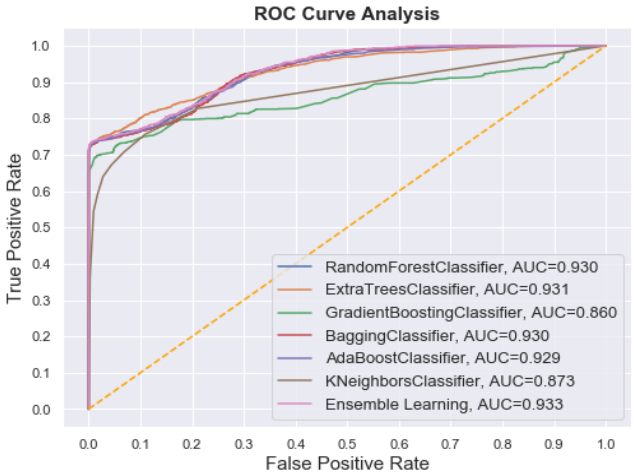

Figure 2: ROC-curves of the supervised learning models trained on Elliptic data set. Area-Under-Curve is denoted by AUC and the bisector straight line is denotes as line of chance.

outperformed also the results provided in the original paper [7], using the same data set as shown in Table 3.

Meanwhile, supervised learning algorithms based decision trees such as Random Forest and Extra Trees methods have revealed remarkable performance which interprets the appropriateness of the used methods on Elliptic data set. k-NN algorithm falls behind all models admitting the least performance in this task with an accuracy equals to 95.1%. Since k-NN is based on Euclidean distances, it is computationally expensive to search for the best values of k. On the other hand, the two drawbacks here of k-NN in our case are the high dimensional of the data set, and the highly imbalanced data set. For example, k-NN relies on the k-neighbours in the feature space to vote for the best class [17]. Due to the existence of numerous negative instances in a neighbourhood of small number of positive instances, the voting mechanism in k-NN is more likely to be skewed toward the majorities.

Table 3: Comparative results between original work in [7] and ours using supervised learning methods on Elliptic data.

Model	Accuracy	F_1 score
Logistic Regression[7]	93.1%	48.1%
Multi-Layer Perceptron[7]	96.2%	65.3%
Random Forest[7]	97.7%	78.8%
Ensemble Learning (our results)	**98.13%**	**83.36%**
Random Forest (our results)	**98.06%**	**82.92%**

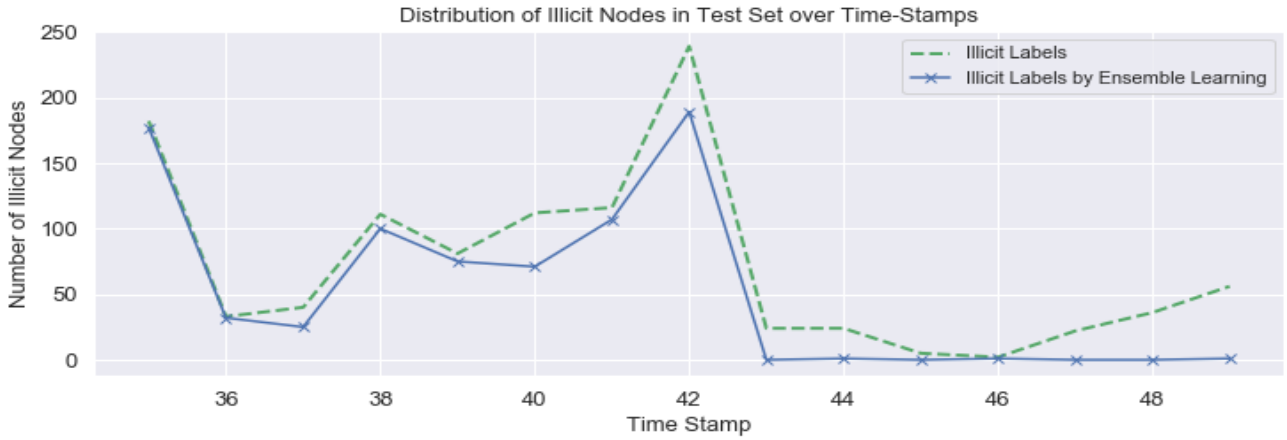

Figure 3: Illicit transactions in test set using true labels and their associated predictions using Ensemble Learning.

That is why Random Forest performed well in this task since it uses a voting mechanism with aggregate the prediction results from a certain number of decision trees, where each tree is trained using a sub-sample of the data set and split of leaves is based on a certain number of features [18]. However, in our experiment, we have chosen to use all the data set to train each tree.

Using ensemble learning, the combination of Random Forest, Extra Trees, and Bagging classifier revealed a potential performance by acquiring the predictions based on averaging the probabilities obtained from these algorithms. Referring to Table 2, ensemble learning has accomplished the minimum number of false positives equal to 7, and thus increasing the precision without significant variation of recall. For instance, the false positive instances might appear commonly between different learning algorithms. These instances will indeed remain the same after using ensemble learning. In contrast, we desire to acquire different classification models, in which each model will fail on classifying correctly certain data points that are distinct in each. Thus, ensemble learning will try to adjust the predicted probabilities from combining several models, so that we can reduce the number of false instances.

From the complexity point of view, Bagging Classifier admits higher time-complexity than Random Forest and Extra Trees, because Bagging here is an ensemble based on Random Forest as a base estimator. Therefore, we associate the time-complexity of ensemble learning method as the complexity of Bagging Classifier, since the latter one describes the worst-case scenario. To do this, let $n =$ training instances, $p =$ number of features, $n_{trees} =$ n_estimators, $d =$ max_depth, n_{voters} = number of constructed voting samples of Bagging Classifier which is set to 10. Thus, the time-complexity of the used ensemble learning, assuming parallel processing of the models, can be expressed as $O(3dn\sqrt{p}.n_{trees}.n_{voters})$. Hence, ensemble learning requires more time-complexity in comparison to the classical methods. But, this will not be matter of interest when dealing with anti-money laundering tasks that require further human intervention.

The maximum AUC is registered for ensemble learning which is equal to 0.933, outperforming the other models as shown in Figure 2. Apparently, the ROC curve of Gradient Boosting Classifier has shown to be the worst performance with respect to the other models. AUC ratio of the latter algorithm is equal to 0.86 which is lower than the k-NN associated with AUC of 0.873.

As shown in Figure 3, the number of illicit transactions at every time-step is plotted for the true labels and the true prediction at these data points of ensemble learning algorithm. Ensemble learning has revealed good discrimination of illicit transactions until 39th time-step. In the range of 40-42, the number of actual illicit transactions have admitted a rapid increase, which after decreased sharply at 43th time-step. This area shows the highest difference between the true-labels and the predicted labels, in which the dark market shutdown occurred referring to [7].

In addition, F_1 scores are plotted for Elliptic data set derived from the performance of the used supervised learning methods as shown in Figure 4. F_1 scores demonstrate the performance of supervised methods regarding the illicit class, which is matter of interest. Not only the performance of ensemble learning has degraded at the dark market shutdown, but also all other supervised methods used in our experiment. As highlighted before, this remarked event has occurred at 43th time-step, where none of the used learning methods was able to detect the illicit transactions. The reason of this degraded performance is due to the occurrence of an event that the algorithm has not learned before.

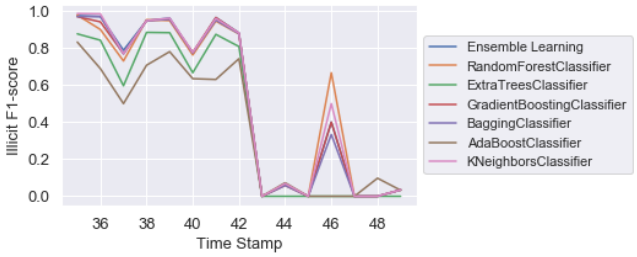

Figure 4: Performance of supervised learning methods on each time-step using Elliptic data set. F_1 score is computed for illicit instances.

As known, Bitcoin graph network of Elliptic data set is a subgraph derived from the transaction graph of Bitcoin blockchain. The other reason that the models are performing poorly on some time-steps can be a result of the loss of structural information derived by the formation of the sub-graph. It is more likely to lose some important links and patterns that are necessary for training the model. As a result, we can assume that the graph network might have similar patterns for both licit and illicit nodes, due to the sub-graph resampling, in which the model is trying to minimize the error on same pattern of different labels. Moreover, the models are not trained on the unknown labels, in which there might be some interesting features and node structures to enhance the detection of

the model. Regarding the high difference between positive and negative instances, it is always desired to have a balanced data set. However, resampling methods such as undersampling and oversampling techniques are not very recommended for this data set. They could decrease the performance or add nothing. Undersampling method tries to reduce the majorities which lead to loss of information for important nodes as well as the edges in the graph structure. On the other hand, oversampling techniques such as Synthetic Minority-Oversampling Technique (SMOTE) will interpolate the data points as well as generating the aggregated features which are not viable. Aggregated features occur from the graph structure by moving one-hop backward forward from the center node, and interpolating such data is a misleading point.

GCN is still an emergent technology, and its applications in graph structures admit promising outcomes. Indeed, GCN requires a convenient data resampling and model tuning to learn the necessary patterns in the graph beside the local information. Nevertheless, the human-intelligence is still needed in anti-money laundering regulations because artificial intelligence arrived to assist not cancel the human-intelligence. The proposed ideas will be raised for future explorations.

6. CONCLUSION AND FUTURE WORK

For the sake of assisting AML processes in Bitcoin via machine learning, we have done a comparative analysis of Elliptic data set to spot illicit transactions using different supervised learning methods. We have shown that the combination of supervised learning methods known as ensemble learning outperformed all other methods using local features and aggregated features derived from Bitcoin transaction graph. Indeed, we aim to reduce the number of false positives without increasing false negatives. The results have shown a noticeable improvement, even though the classical methods outperformed Graph Convolutional Network in the original paper.

In future work, we will explore different supervised learning techniques based on graph-structure along with performing an appropriate pre-processing analysis on the graphs. Indeed, data pre-processing based graph structure is a tricky problem when dealing with graph networks. Future work will take into a consideration graph structure, as well deep learning models in the upcoming approaches.

7. ACKNOWLEDGMENTS

This work is supported by Bournemouth University. Data is publicly available under a Public License thanks to Elliptic company and their domain expertise (www.elliptic.co).

8. REFERENCES

[1] Satoshi Nakamoto et al. 2008. Bitcoin: *A peer-to-peer electronic cash system*. (2008).

[2] Christian Decker and Roger Wattenhofer. 2013. Information propagation in the bitcoin network. In *IEEE P2P 2013 Proceedings*. IEEE, 1–10.

[3] Stefano Bistarelli and Francesco Santini. 2017. Go with the-bitcoin-flow, with visual analytics. In *Proceedings of the 12th International Conference on Availability, Reliability and Security*. 1–6.

[4] Mikkel Alexander Harlev, Haohua Sun Yin, Klaus Christian Langenheldt, Raghava Mukkamala, and Ravi Vatrapu. 2018. Breaking bad: De-anonymising entity types on the bitcoin blockchain using supervised machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

[5] Jesse Bryan Crawford. 2019. Knowing your bitcoin customer: A survey of bitcoin money laundering services and technical solutions for anti-money laundering compliance. (2019).

[6] Malte Möser, Rainer Böhme, and Dominic Breuker.2013. An inquiry into money laundering tools in the Bitcoin ecosystem. In *2013 APWG eCrime Researchers Summit*. Ieee, 1–14.

[7] M. Weber, G. Domeniconi, J. Chen, D. K. I. Weidele, C.Bellei, T. Robinson, and C. E. Leiserson. 2019. Antimoney laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. *KDD Workshop on Anomaly Detection in Finance* (2019).

[8] Nitesh V Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*. Springer, 875–886.

[9] Peter Sollich and Anders Krogh. 1996. Learning with ensembles: How overfitting can be useful. In Advances in neural information processing systems. 190–196.

[10] Issam H Laradji, Mohammad Alshayeb, and Lahouari Ghouti. 2015. Software defect prediction using ensemble learning on selected features. *Information and Software Technology* 58 (2015), 388–402.

[11] J. D. Hunter. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9, 3 (2007), 90–95. DOI: http://dx.doi.org/10.1109/MCSE.2007.55

[12] Thai Pham and Steven Lee. 2016. Anomaly detection in bitcoin network using unsupervised learning methods. Vol. abs/1611.03941.

[13] Thomas N Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

[14] Yi Ma, Jianye Hao, Yaodong Yang, Han Li, Junqi Jin, and Guangyong Chen. 2019. Spectral-based Graph Convolutional Network for Directed Graphs. *arXiv preprint arXiv:1907.08990* (2019).

[15] Martin Harrigan and Christoph Fretter. 2016. The unreasonable effectiveness of address clustering. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. IEEE, 368–373.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B.Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss,V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[17] Min-Ling Zhang and Zhi-Hua Zhou. 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern recognition* 40, 7 (2007), 2038–2048.

[18] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

Authors' background

Your Name	Title*	Research Field	Personal website
Ismail Alarab	Phd candidate	Computer science and technology	
Simant Prakoonwit	Associate professor	AI/ Computer vision	https://staffprofiles.bournemouth.ac.uk/display/sprakoonwit
Mohamed Ikbal Nacer	Phd candidate	Computer science and technology	

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor