31st IEEE International Conference on Computer Theory & Applications

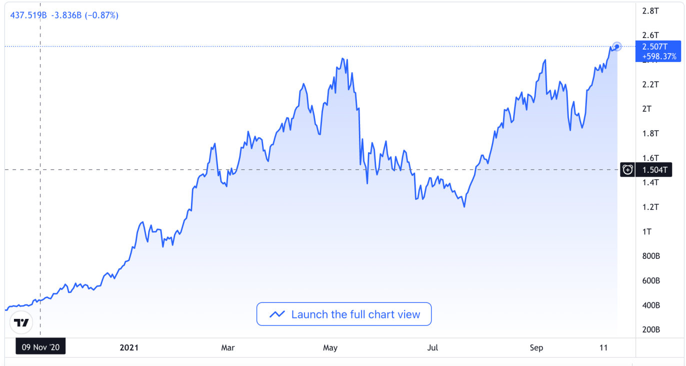
Anomaly Detection on Bitcoin, Ethereum Networks Using GPU-accelerated Machine Learning Methods

Presenting Author

Youssef Elmougy
Georgia Institute of Technology
Georgia, USA
yelmougy3@gatech.edu

Oliver Manzi
Luleå University of Technology
Luleå, Sweden
oliman-8@ltu.se

Cryptocurrencies are becoming prominent

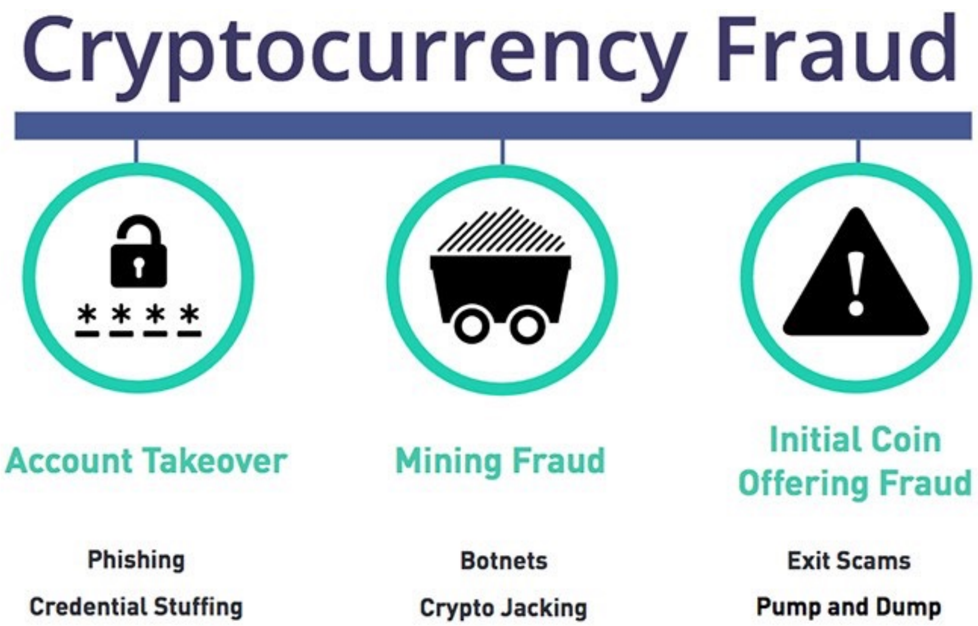


Market cap: $2.3T

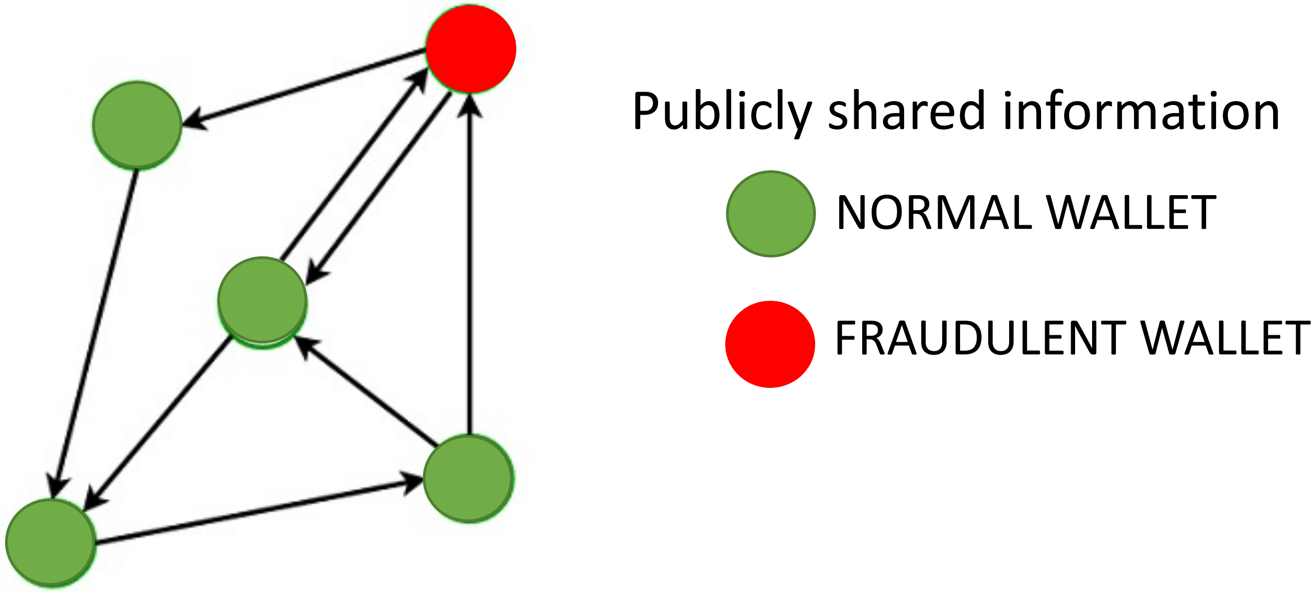
THE BLOCKCHAIN

→ Expand in sectors beyond financial services

Malicious Behavior Towards Fraud Cryptocurrency Holders



2015: Bitcoin-based scams defrauded **13,000** victims and stole over **$11M**



Identifying Fraudulent Transactions and Accounts

Detecting anomalies in the cryptocurrency transaction networks!

Hold 64.5% of market cap

Identifying fraudulent transactions and accounts by detecting anomalies in the Bitcoin and Ethereum transaction networks

Identifying Fraudulent Transactions and Accounts

1. Systematically analyze and accurately detect anomalies on both the Bitcoin and the Ethereum networks
2. Deploy GPU-accelerated ML algorithms to analyze datasets consisting of millions of transactions
3. Show relationships and feature correlation through sensitivity analysis
4. Generalize the study of anomaly detection in the cryptocurrency blockchain to other blockchain networks

Identifying Fraudulent Transactions and Accounts

1. Systematically analyze and accurately detect anomalies on both the Bitcoin and the Ethereum networks
2. Deploy GPU-accelerated ML algorithms to analyze datasets consisting of millions of transactions
3. Show relationships and feature correlation through sensitivity analysis
4. Generalize the study of anomaly detection in the cryptocurrency blockchain to other blockchain networks

Identifying Fraudulent Transactions and Accounts

1. Systematically analyze and accurately detect anomalies on both the Bitcoin and the Ethereum networks
2. Deploy GPU-accelerated ML algorithms to analyze datasets consisting of millions of transactions
3. Show relationships and feature correlation through sensitivity analysis
4. Generalize the study of anomaly detection in the cryptocurrency blockchain to other blockchain networks

Identifying Fraudulent Transactions and Accounts

1. Systematically analyze and accurately detect anomalies on both the Bitcoin and the Ethereum networks
2. Deploy GPU-accelerated ML algorithms to analyze datasets consisting of millions of transactions
3. Show relationships and feature correlation through sensitivity analysis
4. Generalize the study of anomaly detection in the cryptocurrency blockchain to other blockchain networks

Machine Learning Models

Anomaly Transaction Detection

Models trained:

1. Support Vector Machine (SVM)
2. Random Forest
3. Logistic Regression

Machine Learning Models

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$s.t., y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., m$$

Models trained:

1. Support Vector Machine (SVM)
2. Random Forest
3. Logistic Regression

Machine Learning Models

$$C_{RF} \leftarrow majorityVote\{C_i(x)\}_1^n$$

Where $C_i(x)$ is the predicted classification of the i^{th} random tree.

Models trained:

1. Support Vector Machine (SVM)
2. Random Forest
3. Logistic Regression

Machine Learning Models

$$\min_{w,c} \frac{w^T w}{2} + C \sum_{i=1}^{N} \log(\exp(-y_i(x_i^T w + c)) + 1)$$

Where we have a set of d features, $x = (x_1, ..., x_d)$, parameter vector w, and optimal value C calculated via cross validation.

Models trained:

1. Support Vector Machine (SVM)
2. Random Forest
3. Logistic Regression

Dataset Collection and Preparation

Bitcoin Dataset

Total transactions: 30,294,698

Non-fraud: 30,290,045

Fraud: 4,653

Ethereum Dataset

Total accounts: 504,942

Non-fraud: 500,000

Fraud: 4,942

Model Features

Feature	Description
IN-TXS	Number of incoming transactions
OUT-TXS	Number of outgoing transactions
IN-BTC / IN-ETH	Amount (in Bitcoin/Ether) on incoming transactions
OUT-BTC / OUT-ETH	Amount (in Bitcoin/Ether) on outgoing transactions
AVG-IN	Average amount (in Bitcoin/Ether) on incoming transactions
AVG-OUT	Average amount (in Bitcoin/Ether) on outgoing transactions
TOTAL-BTC / TOTAL-ETH	Total amount (in Bitcoin/Ether) on all incoming and outgoing transactions
FRAUD	Fraud boolean classifier

Exploratory Analysis of the Datasets

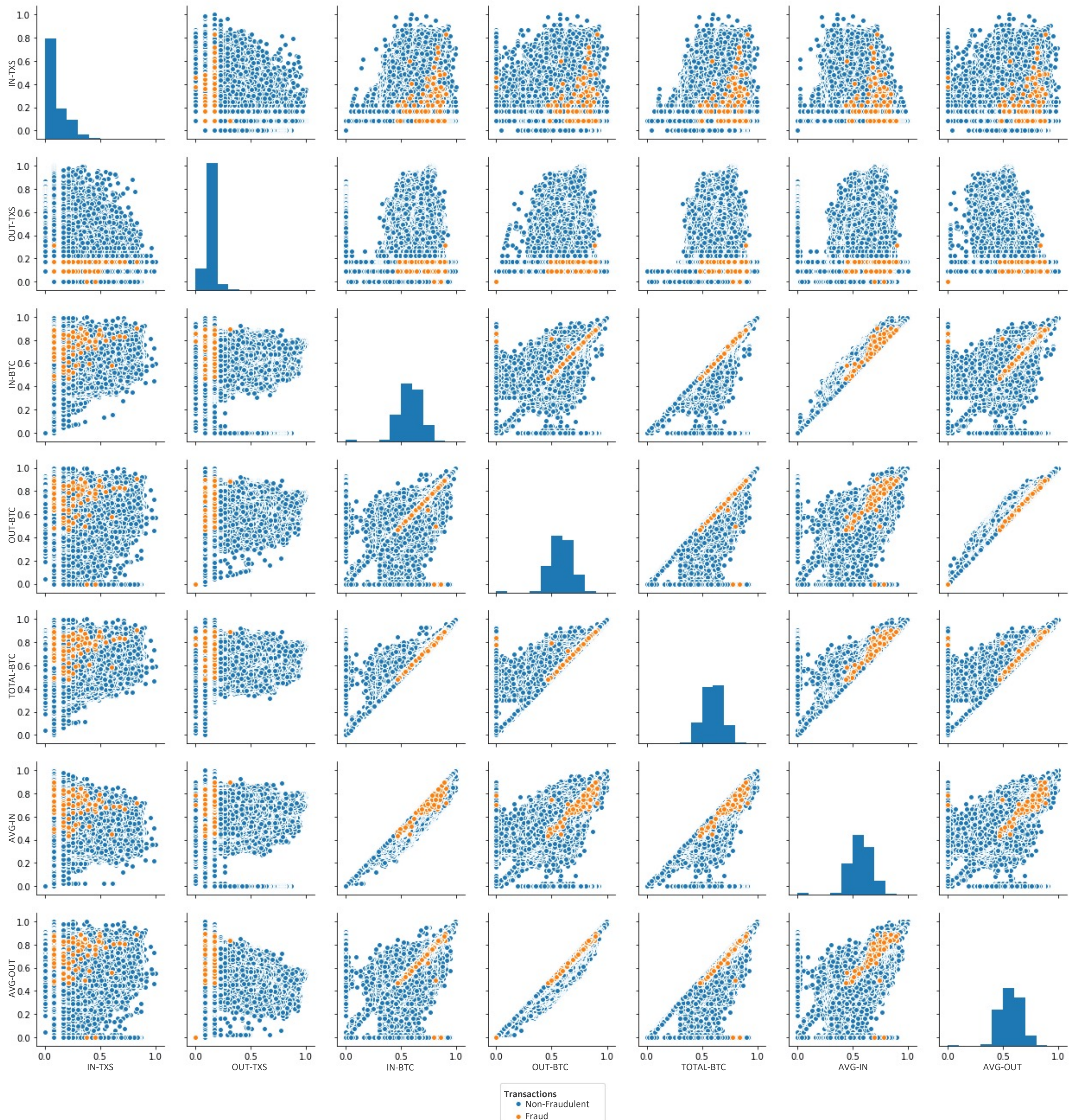
High imbalance of fraudulent and non-fraudulent transactions in both datasets due to low public availability of fraudulent data on the blockchain

Data followed a **right-skewed distribution**

Solved with a $\log(x+1)$ transformation

Normalization and **standardization transformations** were applied

Exploratory Analysis of the Datasets



$\boldsymbol{IN-TXS}$ and $\boldsymbol{OUT-TXS}$ *feature a negative correlation with maliciousness*

$\boldsymbol{IN-BTC, IN-ETH, OUT-BTC, OUT-ETH, AVG-IN, AVG-OUT, TOTAL-BTC}$ *and* $\boldsymbol{TOTAL-ETH}$ *feature a positive correlation with maliciousness*

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

		Actual	
		Non-Fraudulent	Fraudulent
Predicted	Non-Fraudulent	TP	FP
	Fraudulent	FN	TN

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

$$Precision = \frac{tp}{tp + fp}$$

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

$$Recall = \frac{tp}{tp + fn}$$

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

Model Verification

Verification metrics:

1. Confusion Matrix
2. Precision
3. Recall
4. Accuracy
5. F1-Score

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Results of Anomaly Detection

Used an 80/20 split of training and testing

Bitcoin Dataset

Non-fraud: 6,058,009

Fraud: 931

Ethereum Dataset

Non-fraud: 100,000

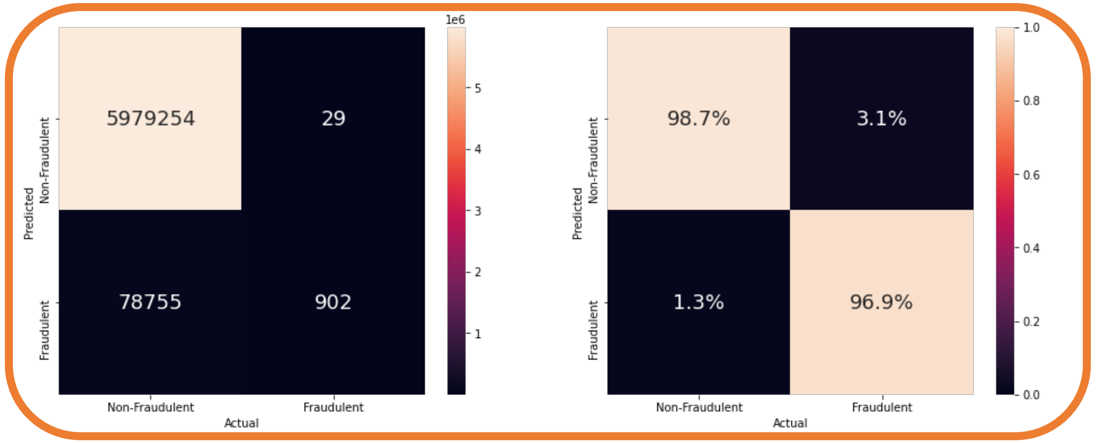
Fraud: 988

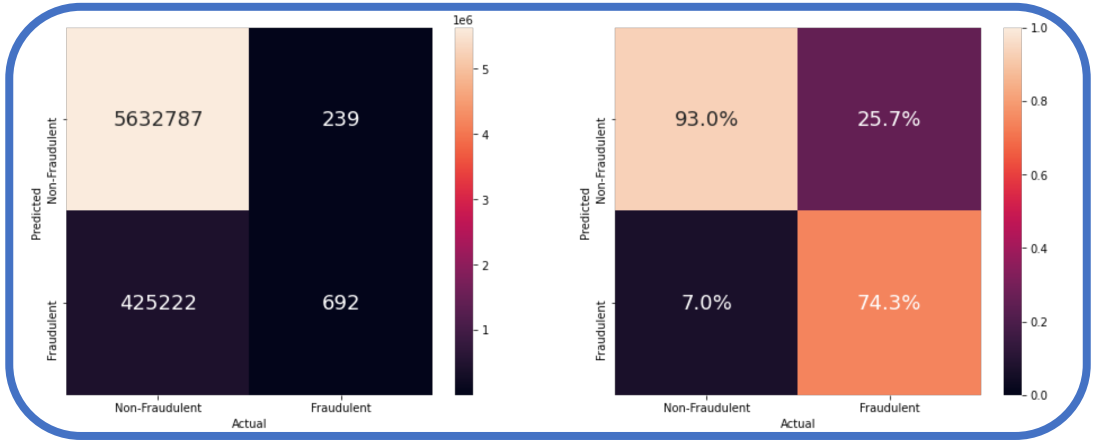
Updated verification metrics:

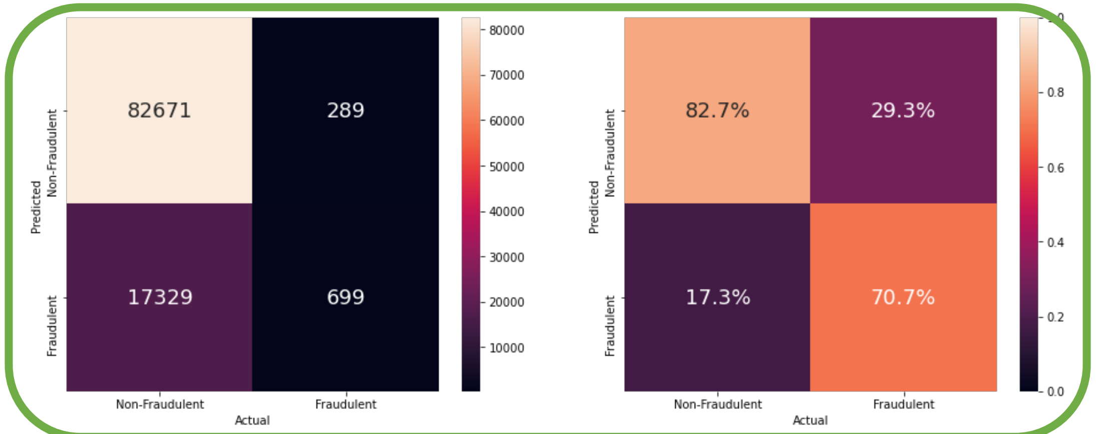
1. Confusion Matrix
2. ~~Precision~~
3. Recall
4. Accuracy
5. F1-Score

Results of Anomaly Detection

Algorithm		Recall	Accuracy	F1 Score
SVM	Bitcoin	0.987	0.987	0.994
	Ethereum	0.827	0.826	0.904
Random Forest	Bitcoin	0.930	0.930	0.964
	Ethereum	0.835	0.834	0.909
Logistic Regression	Bitcoin	0.898	0.897	0.946
	Ethereum	0.819	0.818	0.899







Bitcoin Dataset

SVM

Random Forest

Logistic Regression

Results of Anomaly Detection

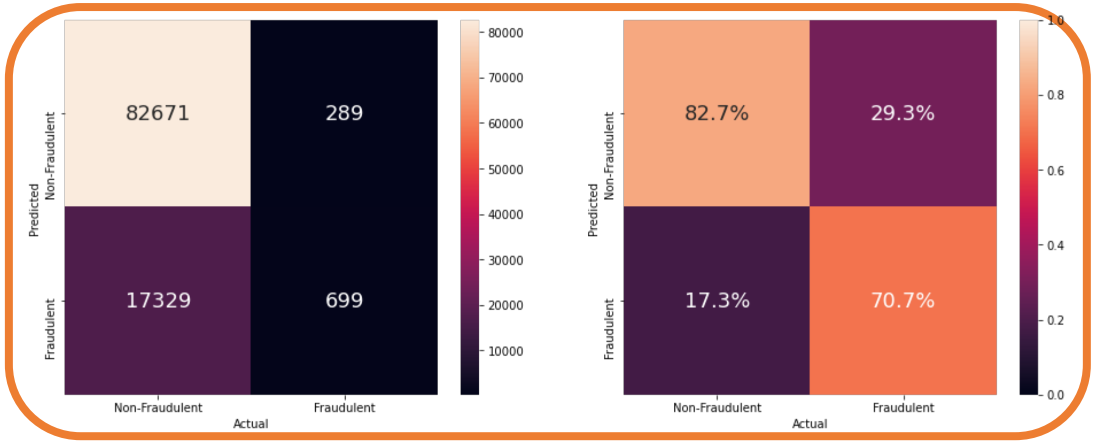
Algorithm		Recall	Accuracy	F1 Score
SVM	Bitcoin	0.987	0.987	0.994
	Ethereum	0.827	0.826	0.904
Random Forest	Bitcoin	0.930	0.930	0.964
	Ethereum	0.835	0.834	0.909
Logistic Regression	Bitcoin	0.898	0.897	0.946
	Ethereum	0.819	0.818	0.899

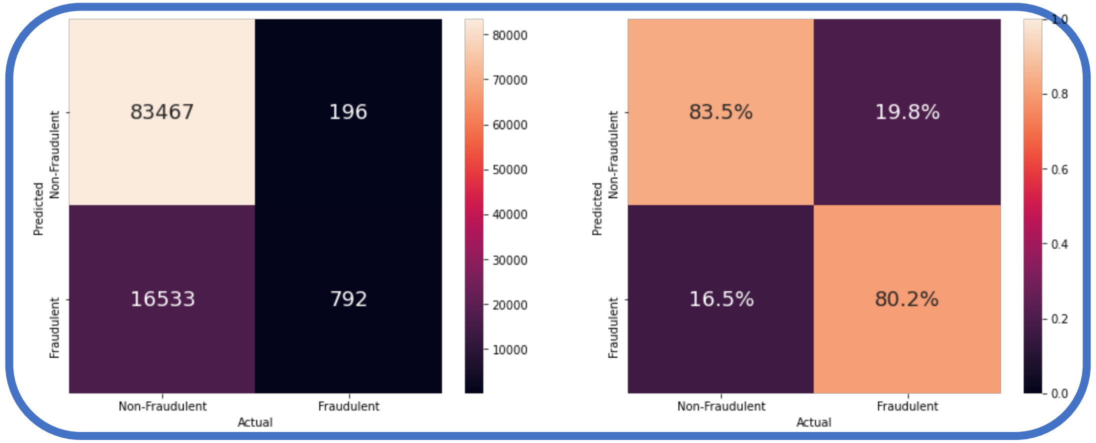
Ethereum Dataset

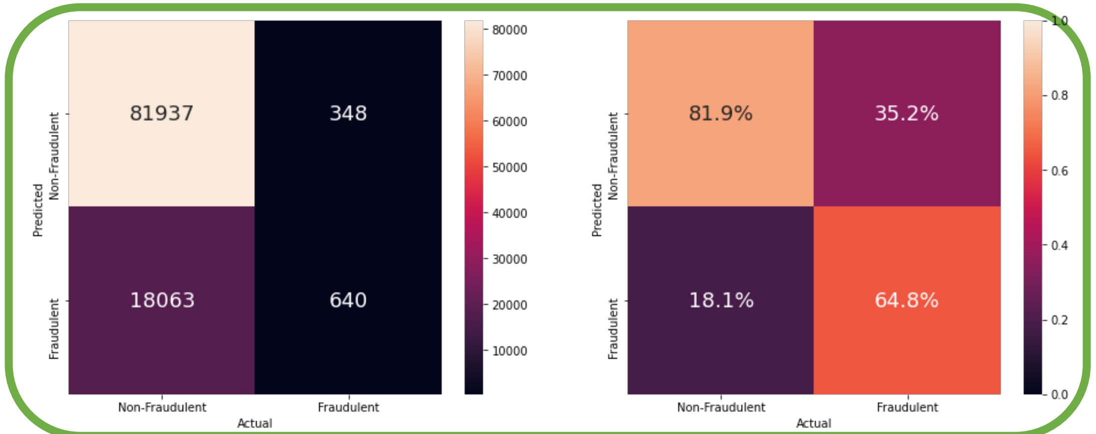
SVM

Random Forest

Logistic Regression







31st IEEE International Conference on Computer Theory & Applications

Thank you for your attention!
Any questions?

Anomaly Detection on Bitcoin, Ethereum Networks Using GPU-accelerated Machine Learning Methods

Youssef Elmougy
Georgia Institute of Technology
Georgia, USA
yelmougy3@gatech.edu

Oliver Manzi
Luleå University of Technology
Luleå, Sweden
oliman-8@ltu.se