Improving Minority-Class Classification in Large, Imbalanced Datasets

By:

Tom Steinman (University of Missouri-Kansas City)

Indraraj Biswas (University of Missouri-Kansas City)

Advisors: Yugyung Lee (Leeyu@umkc.edu)

Luke Miller (ljmbm5@umsystem.edu)

Introduction

Problem

Minority-class classification is difficult in large imbalanced datasets, causing misclassifications in classification models.

Motivation

- In cybersecurity intrusion detection models, many attacks are rare.
- Big datasets have very few examples of these rare, minority classes, so even good models miss them and can lead to blind spots in security monitoring.
- Improving minority classification in cybersecurity is transferable to many domains, including healthcare, medical imaging, insurance, E-commerce, financial investments and more.

Goals

- Improve minority-class detection without hurting other classes.
- Add knowledge graph (KG) context to flows (who talks to whom, how often, how rare).
- Compare methods with and without knowledge graph context.
- Compare classical vs. quantum computing methods.

Dataset & Knowledge Graph

Dataset & Preprocessing

CSE-CIC-IDS2018, (15 classes, 14 attack classes, highly imbalanced) multi-class network flows.

Dataset size post processing: 10189072 Rows, 89 Columns

CSE-CIC-IDS2018 Preproc

Attack Label	Count	%
Benign	9,540,178	93.63%
DDoS attacks-LOIC-HTTP	569,454	5.59%
Infilteration	34,581	0.34%
Bot	24,590	0.24%
FTP-BruteForce	2,526	0.025%
SSH-Bruteforce	5,442	0.053%
DoS attacks-SlowHTTPTest	2,755	0.027%
DoS attacks-Slowloris	2,449	0.024%
DDOS attack-HOIC	1,342	0.01%
DDOS attack-LOIC-UDP	974	0.01%
DoS attacks-GoldenEye	647	0.01%
Brute Force -Web	410	0.004%
DoS attacks-Hulk	198	0.0019 %
Brute Force -XSS	177	0.0017 %
SQL Injection	76	0.0007 %

Knowledge Graph

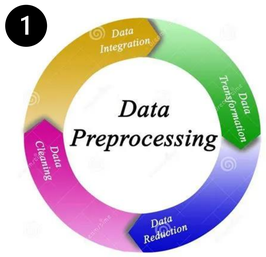
- KG build: For each 5-minute window, make a Host↔Service graph and keep a short rolling history. Also add Host↔Host links when two hosts use the same service (H→S→H).
- KG features: Popularity (degree), rarity (IDF), host–service association (PMI), time spikes (burstiness), and simple metapath counts over time. This data is appended to the dataset.

Quantum Feature Selection

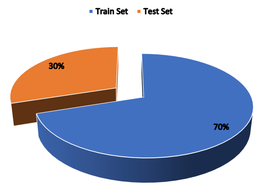
- Quantum search (QUBO → QAOA): the circuit tries different combinations of features, keeping the ones that give the best score (lowest "energy").
- Noise mitigation: we use error-reduction steps to reduce noise (twirling, XY4).

Methodology

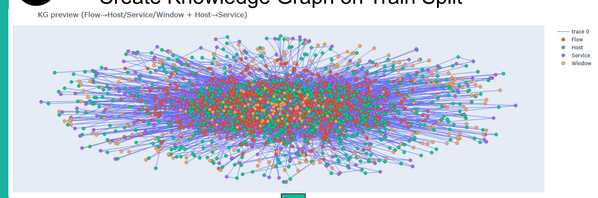
1



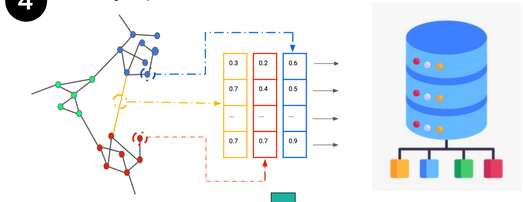
2 70/30 Train/Test Statified Split by time window



3 Create Knowledge Graph on Train Split

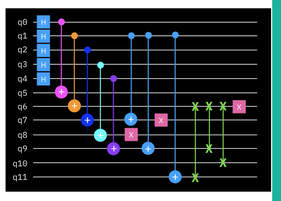


4

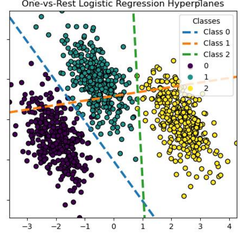


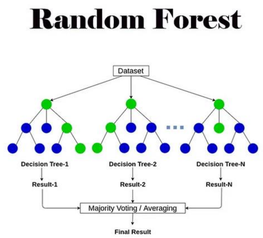
5 Quantum and Classical Feature Selection

- QUBO → Ising Hamiltonian → SparsePauliOp
- Build QAOA circuit (Ansatz)
- Estimator = optimize parameters
- Sampler = sample subsets
- Warm starts
- Read quantum computing errors and mitigate
- Decode and select features

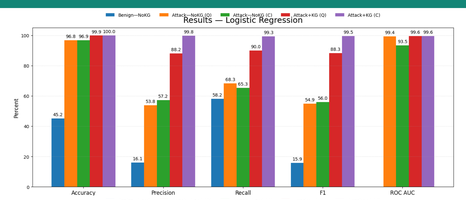


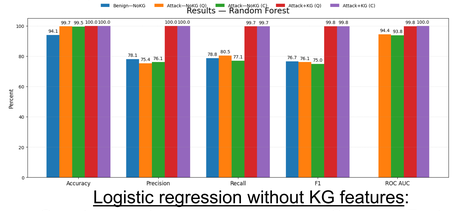
6 Compare Methods: Logistic Regression and Random Forest





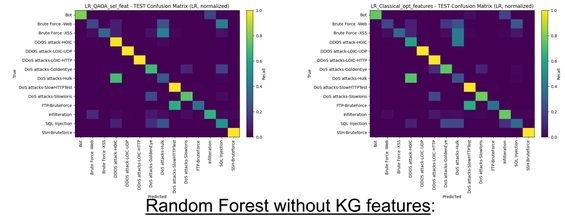
Results





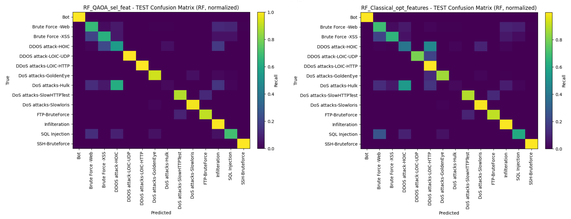
Logistic regression without KG features:

Quantum Features vs Classical Features



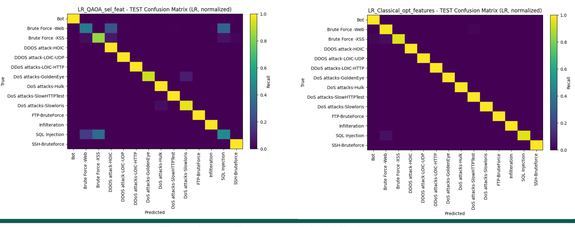
Random Forest without KG features:

Quantum Features vs Classical Features



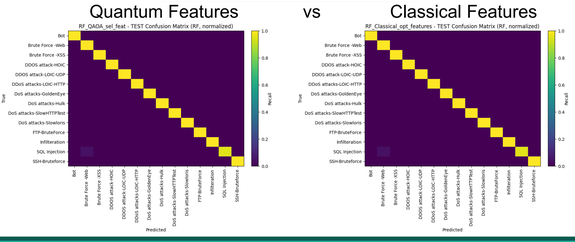
Logistic Regression with KG features:

Quantum Features vs Classical Features



Random Forest with KG features:

Quantum Features vs Classical Features



Conclusion

Comparison (10 Fea

Setup	Model	Selector	Acc %	Prec %	Rec %	F1 %	ROC AUC %
Benign – No KG	LR	Classical	45.2	16.1	58.2	15.9	—
Benign – No KG	RF	Classical	94.1	78.1	78.8	76.7	—
Attack only No KG	LR	Quantum	96.8	53.8	68.3	54.9	99.4
Attack only No KG	LR	Classical	96.9	57.2	65.3	56.0	93.5
Attack only No KG	RF	Quantum	99.7	75.4	80.5	76.1	94.4
Attack only No KG	RF	Classical	99.5	76.1	77.1	75.0	93.8
Attack only with KG	LR	Quantum	99.9	88.2	90.0	88.3	99.6
Attack only with KG	LR	Classical	100	99.8	99.3	99.5	99.6
Attack only with KG	RF	Quantum	100	100	99.7	99.8	99.8
Attack only with KG	RF	Classical	100	100	99.7	99.8	100

KG signals work: Graph and time-aware features improved separation of rare attack types compared to raw flow stats alone.

- Quantum vs. classical selection: The quantum selector chose 3 of 10 KG-derived signals (classical chose 1 of 10), with a 6/10 overall overlap—showing strong agreement while elevating relational and temporal context.
- Quantum selection delivered similar test results to the classical selector, indicating you can keep accuracy while biasing the feature set toward richer KG structure.

Takeaway:

- For imbalanced problems where context matters, **KG + quantum feature selection** is a practical path to better minority-class detection. While quantum metrics were lower, combining features is simple. Quantum feature selection is cheap in comparison to many quantum computing methods. Combining classical and quantum features is an efficient way to improve classical methods.

Why this transfers beyond cybersecurity:

- Healthcare/imaging: Rare subtypes, time correlation
- Insurance: Claim-network links expose organized fraud.
- Finance: Trade patterns; trade anomalies; social media aspects as KG features.
- Any imbalanced domain: KG + quantum boosts rare-event signal.

Future Plans:

- Investigate further quantum methods such as a Quantum Neural Network (QNN) to evaluate efficacy of quantum feature selection.
- Use knowledge graphs as a sanity check for models instead of additional features.
- Apply to other domains.

Acknowledgements

Special thanks to:
Dr. Yugyung Lee and Luke Miller in their guidance on this project.

Link : https://colab.research.google.com/drive/1DC8KBelVWDp3NdKmdVvaZvT0nKKpMgDr?usp=sharing