Towards Robust Malware Classification

Overview

- **Malicious** programs continually evolve over time.
- This causes **challenges** such as **data collection** and **concept drift**.
- We require models to understand when drift has occurred, to adapt to drift, and to make effective use of collected malware samples.

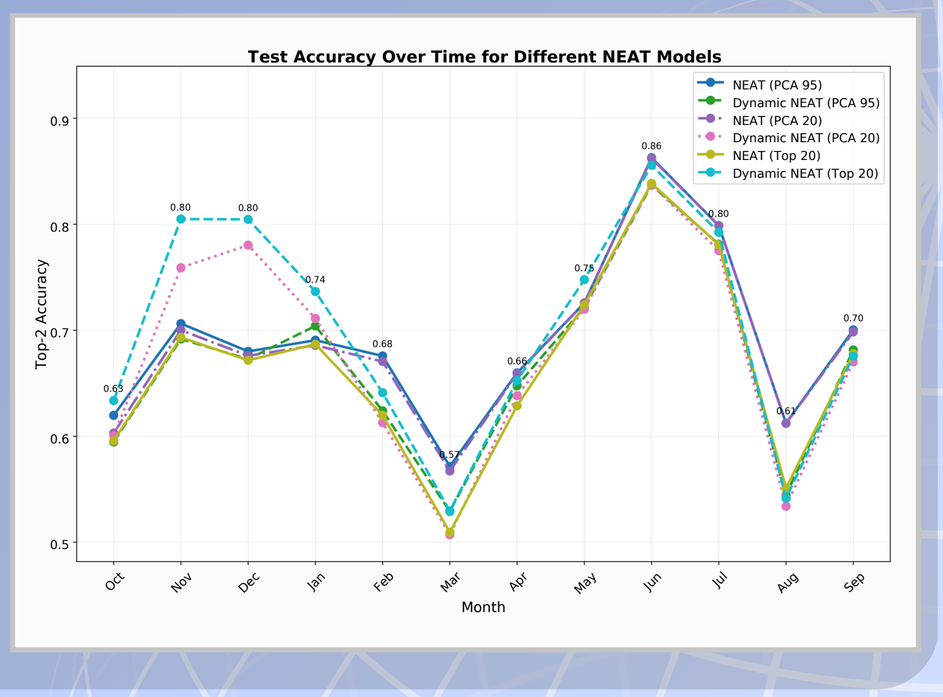
Objectives

1) Adaptive Malware Classification
2) Robust Data Augmentation
3) Concept Drift Explainability

Adaptive Topology Methods

Method & Results

- We trained **NEAT** model variants for **malware classification**.
- **DynNEAT** achieved highest **adaptive** topology method accuracy of **65.78%**.



Conclusion

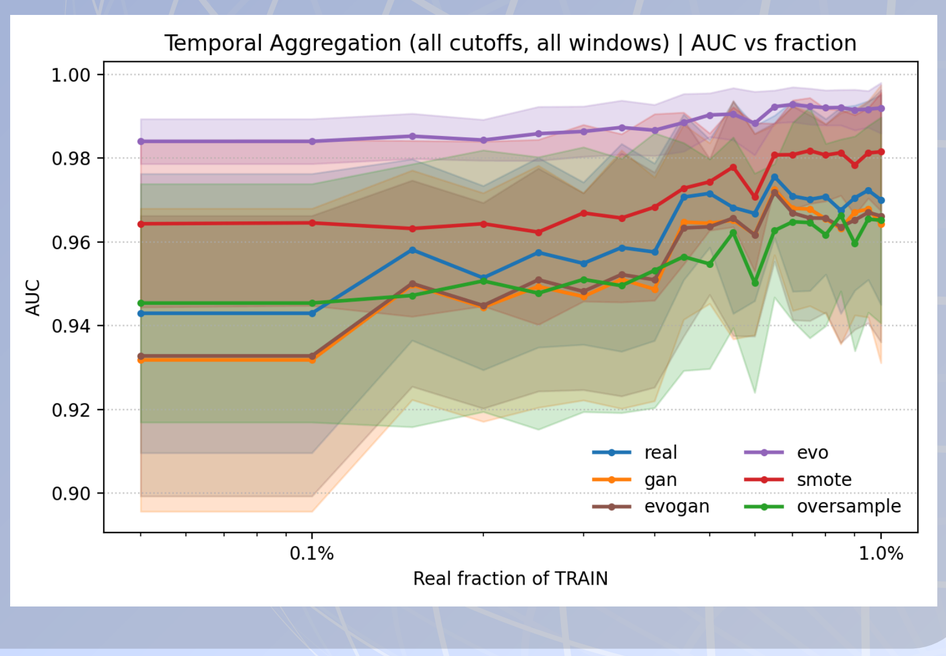
- Fixed topology outperforms adaptive topology methods.
- **Dimensionality** impacts learning with NEAT models.

Research	Accuracy	Type	Number of classes	Temporal analysis
Lu et al. [16]	96.96%	Multi	11	No
Lu et al. [16]	93.42%	Multi	49	No
Louk & Tama [15]	99.97%	Binary	-	No
Guldemir et. al [8]	89.55%	Multi	20	Yes
Our work	65.78%	Multi	20	Yes

Synthetic Data Augmentation

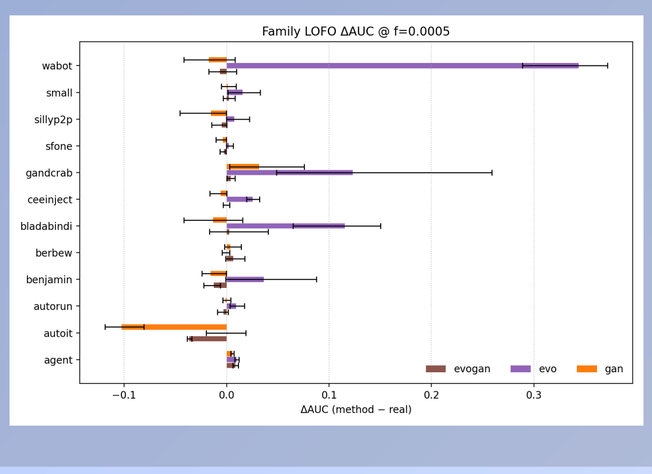
Methods

- Synthesising malware positives with **WGAN-GP**, **EVO**, and **EvoGAN** to augment scarce training data and boost detection
- Evaluating **i.i.d.**, **temporal**, and **family** regimes using Random Forest



Conclusion

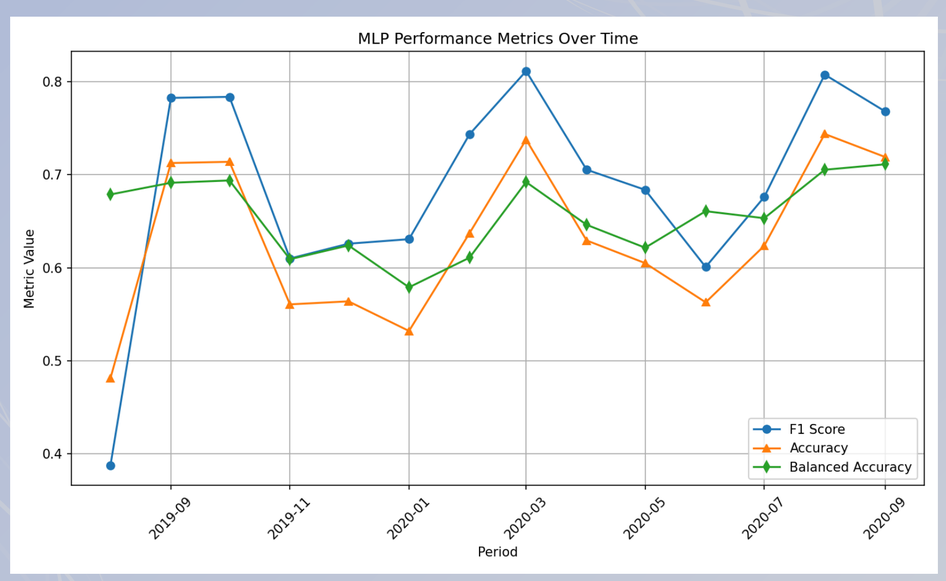
- **EVO outperforms** all baselines and other augmentation methods on **ROC–AUC** across regimes, with the lowest seed variability.
- **Gains** are largest under **scarcity** and taper as real data approaches **~1%**.



Concept Drift Monitoring

Methods

- Training DL models on **historical** malware and **temporally** testing on newer malware to quantify the effects of concept drift.
- Evaluating drift detection performance and providing **insights** on drift through **Explainable AI**.



Conclusion

- The MLP model showed significant performance **degradation**; unexpectedly due to benign samples.
- Ensemble drift detection is best suited to balance **reliability** and **responsiveness.**
- Explainability reports gave insight on feature **distribution shifts** driving concept drift.

University of Cape Town
SCHOOL OF IT
University of Cape Town
Department of Computer Science
www.sit.uct.ac.za

Shaylin Velen <VLNSHA004@myuct.ac.za>
Callum Musselwhite <MSSCAL002@myuct.ac.za>
Simphile Mkhize <MKHSIM067@myuct.ac.za>
Supervised by **Geoff Nitschke <geoff.nitschke@uct.ac.za>**