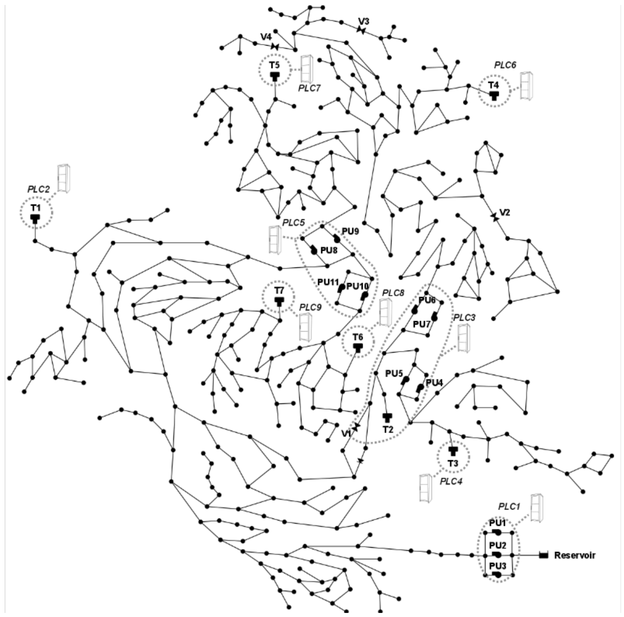
IDENTIFYING CYBER ATTACKS IN WATER DISTRIBUTION SYSTEMS USING AI MODELS

Mehna Lakshminarayanan and Md. Rakibul Ahasan
FSU Cyber-Physical Machine Learning Lab

Introduction

On February 5, 2021, an unauthorized remote access event at a Florida water treatment plant, serving 15,000 residents, resulted in the sodium hydroxide concentration being increased nearly 100-fold; to deathly levels. The operator immediately corrected this, preventing perilous amounts of sodium hydroxide levels in drinking water, but this event underscores the cybersecurity risks of digitization in water distribution centers (WDS).

- denial of service (DoS)
  - disrupts communication between sensors
- replay attacks
  - valid data is fraudulently repeated or delayed
- data manipulation
  - alters sensor readings



Objective

A solution to current problems in WDS is using machine learning (ML) to predict attacks prematurely by using three classifiers: Random Forest, Recurrent Neural Network, and Gated Recurrent Unit. To address the multi-dimensional characteristics preprocessing included variance filtering and downsampling for class imbalances.

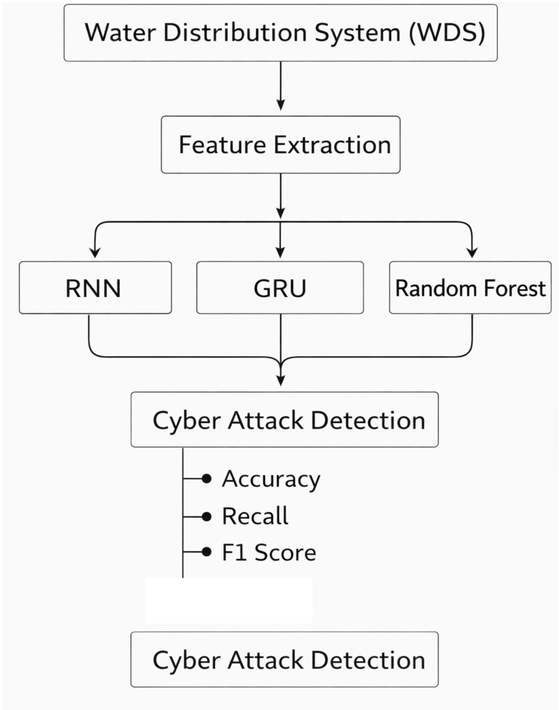
Contributions

FEATURE EXTRACTION

- Features extracted from C-Town water distribution system simulation data
  - Readings: pressure, head, and demand measurements
- Timestamps used: 1–300 for training and 301–425 for testing
- Oversampling + class weighting
  - Sequence-level oversampling
  - Stratified temporal splitting
- Performance evaluated using confusion matrices to compute detection rate, F1 score, and accuracy
- False alarm rates (FAR) monitored as a secondary constraint on model selection

TUNING

- RNN (Baseline SimpleRNN → Improved SimpleRNN):
  - Threshold tuning via on validation ROC curve
- Random Forest (Baseline → Feature-Optimized RF):
  - SelectKBest feature selection with VarianceThreshold pre-filtering
- GRU (Baseline → Deep GRU):
  - Focal Loss (γ=3, α=0.85) to focus training on attack samples
  - Deeper architecture (64→32 units)



Analysis

- Temporal Signatures
  - Deep learning models, especially GRU, captured sequential attack patterns, demonstrating the importance of temporal feature modeling in WDS cyberattack detection
- Class Imbalances
  - Models produced high false positives
  - Suggests further refinements in imbalance-handling strategies are needed
- Threshold Optimization Trade-offs
  - Youden's J statistic improved recall but amplified FAR
  - Illustrates tension between maximizing attack detection and minimizing false alarms in operational WDS contexts

Model	Metric	Start Score	End Score
RF	F1	0.0364	0.5994
	Accuracy	0.2759	0.7689
	Detection Rate	0.0293	0.4345
RNN	F1	0.1964	0.7708
	Accuracy	0.2105	0.8479
	Detection Rate	0.1774	0.8774
GRU	F1	0.1091	0.6154
	Accuracy	0.1404	0.4737
	Detection Rate	0.0952	0.7619

References

Conclusion

DISCUSSION

GRU and SimpleRNN effectively captured temporal attack patterns, which shows strong improvements in F1 and recall. RF demonstrated detection performance without sequential modeling. However, maximizing recall increased false alarms across all three architectures. It underscores the trade-off between attack detection sensitivity and false alarm minimization.

LIMITATIONS

- Class imbalances produced high FAR despite oversampling and class weighting across all models
- Junction features confused relevant patterns even after variance filtering and SelectKBest selection
- GRU showed low AUC (0.17)
- Results are specific to the C-Town simulation dataset and may not generalize to real-world or alternative WDS configurations

IMPLICATIONS

- ML models detect early cyberattacks in water distribution systems
- Using feature statistics and threshold calibration are key to balancing recall and false alarms
- These methods extend to other cyber-physical
- Future work could explore domain-specific matters and transfer learning across WDS topologies