Key Research Challenges and Solutions

Challenge 1: Model Performance vs. Interpretability

- Random Forest achieved best performance (90.4% accuracy, 98.7% precision)
- Ensemble of hundreds of trees difficult to explain to security analysts
- Analysts need actionable intelligence, not just binary predictions

Outcome: Two-stage pipeline provides high accuracy with actionable threat intelligence

Challenge 2: Feature Engineering & Precision-Recall Tradeoff

- Baseline DBSCAN: 93.9% recall, 3.06% false positive rate (excellent balance)
- TF-IDF enhancement: Improved recall to 99.9% but degraded FPR to 41.8%
- More complex features do not guarantee better real-world performance

Outcome: Simpler baseline model better suited for production deployment

Challenge 3: Hyperparameter Tuning Efficiency

- XGBoost has 12+ tunable parameters; exhaustive search requires 48 hours
- Staged tuning approach: Sequential optimization of learning rate, depth, regularization
- Achieved near-optimal performance in 3 hours (16× faster than grid search)

Outcome: Systematic staged approach beats brute force for complex model tuning