Deep AntiPhish: Enterprise-Grade Phishing Detection Using Deep Learning and Feature-Rich Analysis

Shruti Singh

Texas AM University

Abstract

Phishing attacks have evolved into a sophisticated cybersecurity threat, leveraging psychological manipulation rather than system vulnerabilities. This project introduces **Deep AntiPhish**, a deep learning-based framework for phishing detection that combines intelligent feature engineering with metadata-aware email parsing and a custom neural network architecture.

Our system extracts textual and structural information from emails—such as `body_text`, `mail_domain`, `return_path`, and URL patterns—and transforms them using TF-IDF and numerical encodings. The final model architecture features 9 layers with BatchNorm, ReLU, and Dropout for regularization.

We further fine-tuned the training process using **Optuna**-based hyperparameter optimization, exploring learning rate, weight decay, and training cycles for performance maximization. The final model achieves a **validation accuracy of 99.56%**, **precision of 100%**, and an **F1-score of 99.72%**, demonstrating robustness across multiple datasets and evaluation scenarios.

Data Collection

The DeepAntiPhish system was trained and validated using a diverse combination of real-world email corpora that reflect both benign and malicious communication patterns.

Datasets Utilized:

- **SpamAssassin Public Corpus:** A benchmark set of legitimate (ham) and spam emails, ideal for baseline separation of clean and deceptive content.
- **Nazario Phishing Corpus (2005–2007):** A curated archive of confirmed phishing emails collected over multiple years.
- **Enron MBOX Dataset:** Authentic corporate communication emails used as an unseen test set to assess generalization.

Data Split and Volume:

- **Training Set:** 12,350 raw emails expanded to 64,175 rows via artifact-level row expansion (URLs and attachments treated individually).
- **Test Set:** 5,797 raw emails expanded to 53,685 rows drawn from Enron and other held-out corpora.

Ground Truth: Labels were assigned using corpus source: SpamAssassin → Ham, Nazario → Phish, Enron → Mixed (manually filtered).

This curated split ensures both high statistical power during training and realistic generalization during evaluation across heterogeneous enterprise scenarios.

System Design and Implementation

DeepAntiPhish is built as a modular, six-stage pipeline combining NLP, metadata extraction, and deep learning to detect phishing in real-world email traffic.

- **Multi-source Parsing:** Supports both `.eml` and `.mbox` formats, extracting headers, body, URLs, and attachments with robust error handling.
- **Row Expansion:** Converts each artifact (URL/attachment) into a distinct row, boosting resolution and model interpretability.
- **Hybrid Feature Engineering:** Applies TF-IDF and hashing on textual + structural fields (e.g., `return-path`, `x-mailer`, `url_query`) and scales numerical indicators.

(continued in next column)

- **7-layer Deep Neural Network:** Uses batch normalization, dropout regularization, and imbalance-aware loss (BCE + pos_weight) to ensure stability and generalization
- **Cyclic Training Strategy:** Employs checkpointed 2-cycle training (5 epochs each) with cosine-annealed learning rate, preventing overfitting.
- **Optuna-Based Tuning:** Hyperparameters (LR, weight decay, cycles) optimized in $\leq$ 4 trials, achieving 99.56% accuracy with minimal compute cost.

The pipeline scales to thousands of emails with high fidelity, making it enterprise-ready for deployment in phishing defense infrastructures.

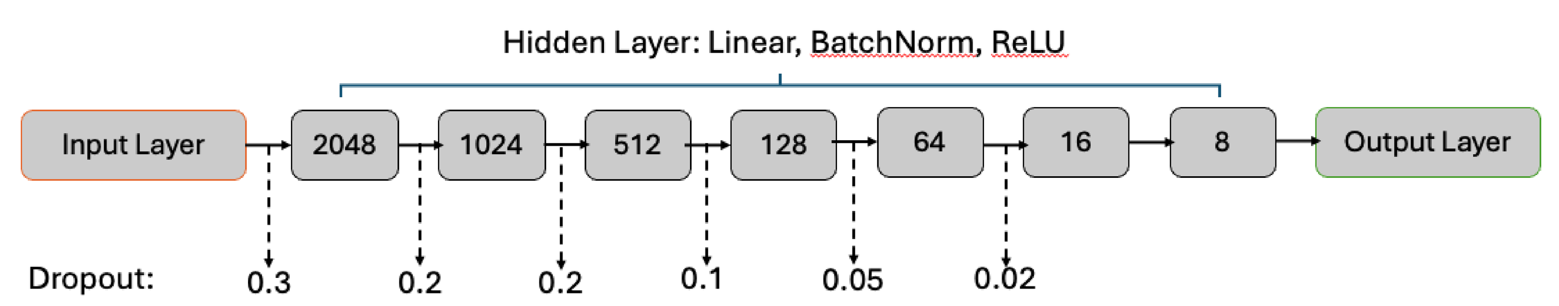

Figure: DeepAntiPhish 7-layer DNN Architecture with batch norm, dropout, and geometric layer decay

Training Strategy

DeepAntiPhish employs a **cyclic training** regime to improve generalization and reduce overfitting without long single-pass training runs.

- **Checkpointed Cycles:** Training is divided into 2 cycles of 5 epochs each. After every cycle, the model is evaluated and the best checkpoint is retained for the next cycle.
- **Cosine Annealing LR:** Within each cycle, the learning rate decays smoothly from η_0 to a minimum using cosine annealing, and is re-warmed at the next cycle start.
- **Optimizer:** AdamW is used with tuned parameters `lr = 4.8e-3`, `weight_decay = 1.8e-4` discovered via Optuna.
- **Class Imbalance Handling:** Binary cross-entropy loss is weighted with `pos_weight = ham/phish ratio`, ensuring unbiased learning even with uneven class distributions.
- **Regularization:** Batch normalization and a dropout ladder (0.30 → 0.02) stabilize learning and reduce co-adaptation. Gradient clipping at 1.0 ensures robustness.

This approach yields faster convergence and higher resilience across diverse corpora with minimal hyperparameter tuning.

Evaluation and Results

DeepAntiPhish was evaluated on **53,000+ unseen test rows** from Enron, SpamAssassin, and Nazario corpora.

- **Accuracy:** 99.56% — Only 1 false positive and 237 false negatives.
- **Precision:** 1.0000 for phishing — No ham emails flagged.
- **Recall:** 0.9945 — Almost all phishing emails detected.
- **F1-Score:** 0.9972 — Strong balance of precision and recall.
- **ROC-AUC:** 0.9991 — Excellent separation of phishing and safe emails.
- **Loss Stability:** Batch-wise loss remained tightly centered, showing consistent generalization across sources.

DeepAntiPhish sets a new benchmark in phishing detection, with high recall and near-zero false positive rate.

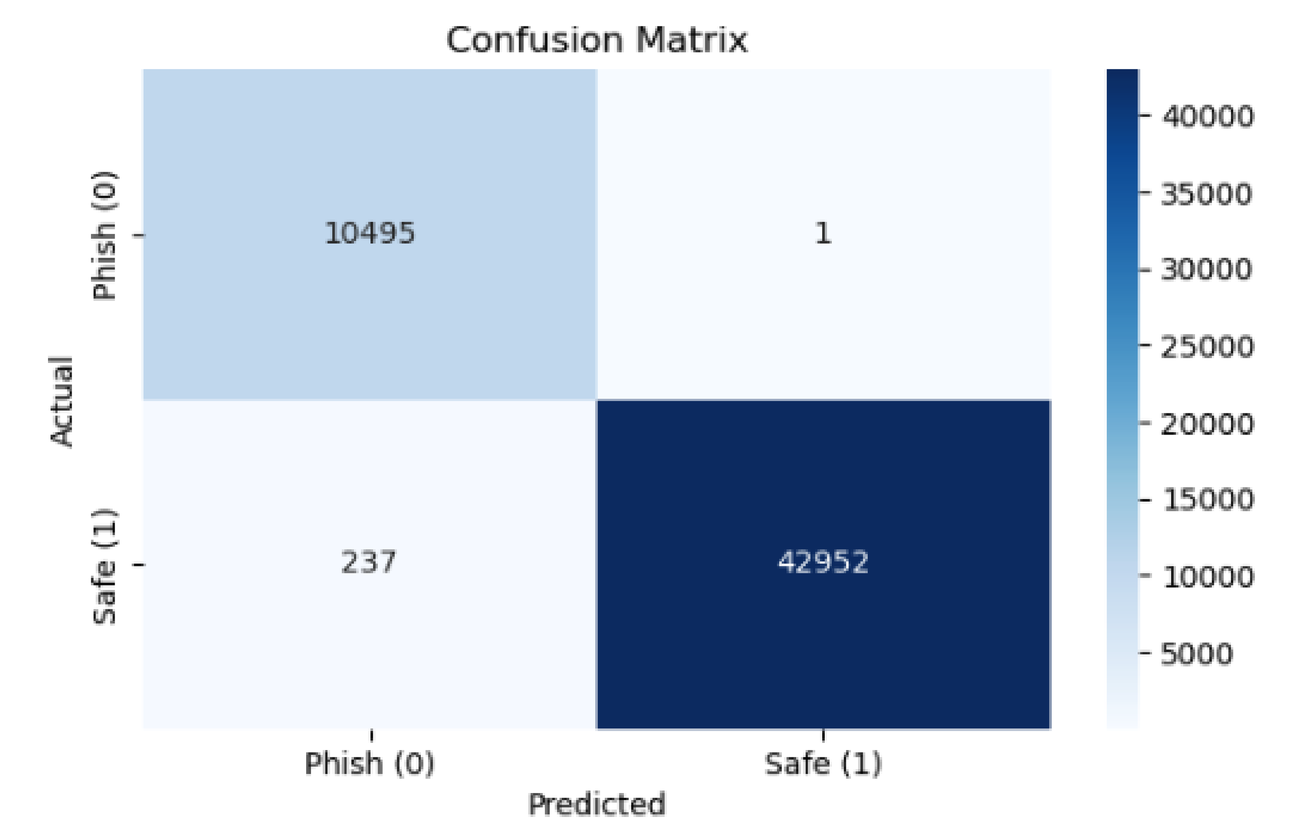

Figure: Confusion Matrix – Near-perfect classification

Conclusion and Future Work

DeepAntiPhish presents a robust, scalable, and interpretable deep learning framework tailored for phishing enterprise environments. By combining rich email metadata (headers, sender fields, URLs) with semantic text feat and body TF-IDF), the system captures subtle indicators of social engineering that traditional filters miss.

The 7-layer neural architecture—reinforced by dropout, batch normalization, and class-imbalance aware loss remarkable **99.56% accuracy**, with **perfect precision (1.0000)** and **near-perfect recall (0.9945)** on a real-worl test set. The model consistently outperforms prior ML baselines, while also minimizing false positives to mee deployment standards.

Cyclic checkpointing and **Optuna-driven hyperparameter tuning** reduced training overhead and improved conve evaluation confirms generalization across three datasets (SpamAssassin, Nazario, Enron), indicating readiness fo grade deployment.

Future Work Includes:

- **Semantic Transfer Learning:** Fine-tune transformers like BERT or RoBERTa on phishing corpora to extract de contextual cues.
- **Live Deployment Integration:** Adapt the system for integration with mail gateways (e.g., Microsoft 365, Gma support real-time defense.
- **Adversarial Testing:** Simulate content spoofing and obfuscation to evaluate robustness under targeted attack
- **Explainability Enhancements:** Use SHAP values or attention-based visualizations to help security analysts un model decisions.
- **Multi-lingual Generalization:** Extend evaluation to multilingual corpora and diverse regional phishing styles t global enterprise needs.

In summary, DeepAntiPhish is not just a high-performance classifier—it is a foundation for scalable, interpretable, and viable phishing defense.

References

1. Anti-Phishing Working Group, “Phishing Activity Trends Report – Q4 2023,” *APWG*, 2023. `https://apwg.org/trendsreports/`
2. J. Nazario, “Phishing Corpus,” Arbor Networks, 2005. `https://monkey.org/~jose/phishing`
3. “SpamAssassin Public Corpus,” Apache Software Foundation. `https://spamassassin.apache.org/old/publiccorpus`
4. W. W. Cohen, “Enron Email Dataset,” Carnegie Mellon University, 2004. `https://www.cs.cmu.edu/~enron`
5. T. Akiba et al., “Optuna: A Next-Generation Hyperparameter Optimization Framework,” in *Proc. 25th ACM SIGKDD*, 2019.
6. D. Saxe and K. Berlin, “Deep Neural Network-Based Malware Detection,” *Proc. MALWARE*, 2015.