# From Ensemble and DistilBERT Baselines to Deep Metric Learning: An Extended Study for Imbalanced Windows PE Malware Detection

Ly Ngoc Vu
*Industrial University of Ho Chi Minh City*
Ho Chi Minh City, Vietnam
dezzhuge@gmail.com

Dang Thi Phuc
*Industrial University of Ho Chi Minh City*
Ho Chi Minh City, Vietnam
phucdt@iuh.edu.vn

***Abstract*—This paper extends a prior ATC 2024 study on Windows PE malware detection, which compared machine learning ensembles and text-based deep learning models (LSTM, BiLSTM, DistilBERT). The previous version achieved strong performance near 99% accuracy on a large PE dataset. In this extended version, we add a deep metric learning (DML) branch on a unified Transformer backbone and evaluate five variants: baseline classification, ArcFace, Contrastive, Triplet, and Multi-Similarity. We report multi-seed statistics and low-false-positive operating behavior (TPR@FPR) for deployment-oriented analysis. Results show that DML is useful but objective-dependent: Multi-Similarity is strongest overall, while ArcFace is unstable under strict low-FPR operation.**

***Index Terms*—malware detection, PE files, deep metric learning, Transformer, low-FPR evaluation**

## I. Introduction

Malware detection on Windows PE files remains challenging under severe class imbalance and strict false-positive requirements. Prior experiments using classical ML and text-based deep learning showed strong aggregate accuracy. However, deployment effectiveness depends on behavior at strict operating points, not only global metrics.

This extension introduces deep metric learning (DML) objectives and compares them under a unified pipeline to assess both classification quality and low-FPR sensitivity.

## II. Relationship to Prior Work

### A. Prior ATC 2024 Results

The prior paper reported:

1) Classical ML models: Logistic Regression, Random Forest, SVC, XGBoost, plus ensemble voting/stacking.
2) Text-based deep learning models: LSTM, BiLSTM, DistilBERT.
3) Large PE dataset setting (˜34k samples) with high performance.

### B. Extension in This Paper

This work adds:

1) Objective-level study of DML variants.
2) Multi-seed statistical evaluation.
3) Deployment-oriented low-FPR analysis.

## III. Methodology

A shared Transformer backbone is used, while objectives/heads vary:

1) baseline: cross-entropy/focal classification.
2) arcface: angular margin objective.
3) contrastive: pairwise distance objective + classification.
4) triplet: relative distance objective with hard mining.
5) multi_similarity: weighted hard-pair similarity objective.

### A. Pipeline Overview

## IV. Results

### A. Prior ATC 2024 Baselines

TABLE I
ML BASELINES FROM PRIOR PAPER

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| Logistic Regression | 0.990086 | 0.990232 | 0.990111 | 0.990112 |
| Random Forest | 0.990000 | 0.990382 | 0.990402 | 0.990403 |
| SVC | 0.990000 | 0.988060 | 0.988075 | 0.988076 |
| XGBoost | 0.990000 | 0.991542 | 0.991566 | 0.991566 |

TABLE II
DEEP LEARNING BASELINES FROM PRIOR PAPER

| Model | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|
| DistilBERT | 0.9864895 | 0.9865220 | 0.9864705 | 0.986471 |
| LSTM | 0.9674135 | 0.9674490 | 0.9674130 | 0.967413 |
| BiLSTM | 0.9507005 | 0.9499545 | 0.9500705 | 0.950102 |

### B. New DML Results

## V. Discussion

Main findings:

1) DML helps, but improvements are objective-dependent.

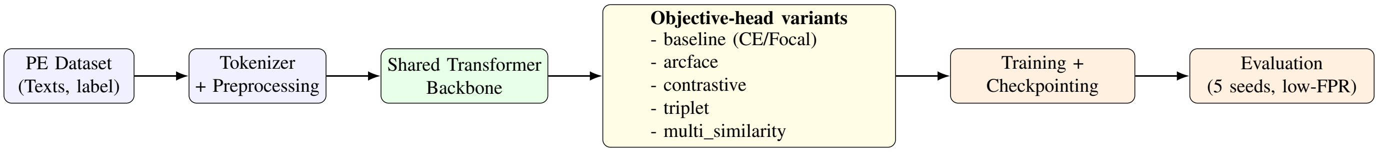

Fig. 1. Unified pipeline for baseline and DML variants.

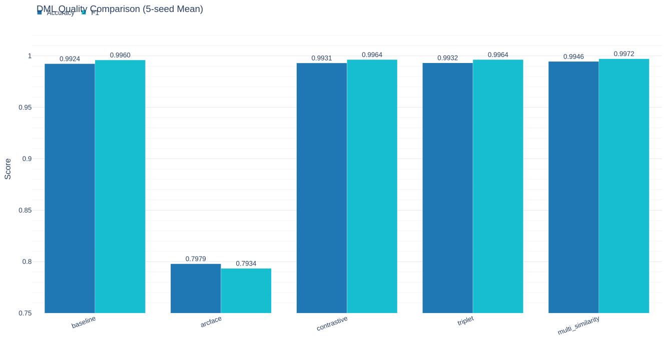

Fig. 2. Accuracy and F1 across DML variants (5-seed mean).

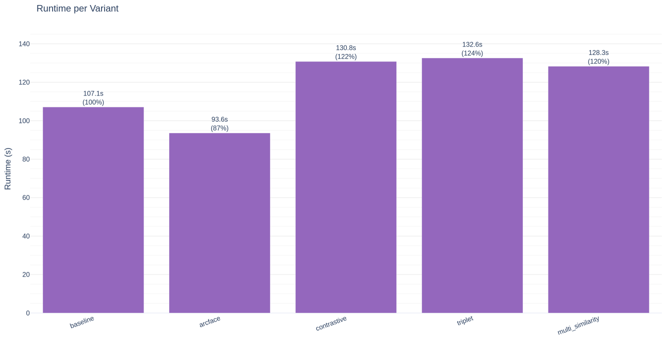

Fig. 3. Single-run runtime per variant.

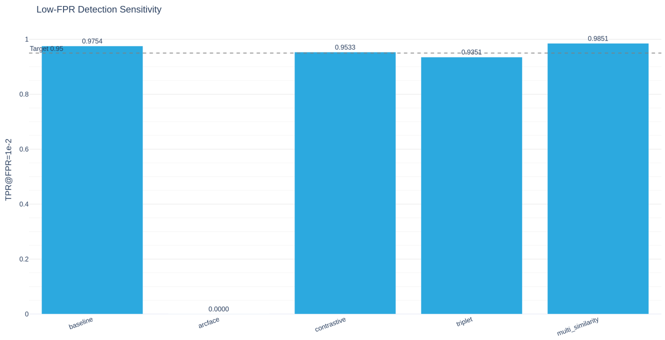

Fig. 4. Low-FPR sensitivity comparison; ArcFace is unstable at strict operating point.

TABLE III
AGGREGATED 5-SEED METRICS (MEAN)

| Variant | Accuracy | F1 | ROC-AUC | PR-AUC | TPR@FPR=1e-2 |
|---|---|---|---|---|---|
| baseline | 0.9924 | 0.9960 | 0.9983 | 0.9999 | 0.9754 |
| arcface | 0.7979 | 0.7934 | 0.9704 | 0.9984 | 0.0000 |
| contrastive | 0.9931 | 0.9964 | 0.9971 | 0.9997 | 0.9533 |
| triplet | 0.9932 | 0.9964 | 0.9969 | 0.9998 | 0.9351 |
| multi_similarity | **0.9946** | **0.9972** | 0.9978 | 0.9999 | **0.9851** |

2) Multi-Similarity is strongest overall in this setup.
3) Baseline remains highly competitive.
4) ArcFace requires further calibration/tuning for strict low-FPR deployment.

## VI. CONCLUSION

This extended study confirms that text-based PE malware detection can achieve high performance and that DML can further improve deployment-oriented behavior when the objective is carefully selected. Multi-Similarity is the recommended candidate in the current experiments.

## ACKNOWLEDGMENT

This work extends the authors' prior ATC 2024 study and integrates new metric-learning experiments and analysis.

## REFERENCES

[1] L. N. Vu and D. T. Phuc, ”Windows Malware Detection: Exploring from Machine Learning to Text-Based Deep Learning Approaches,” in Proc. ATC, 2024.

[2] J. Deng et al., ”ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” CVPR, 2019.

[3] X. Wang et al., ”Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning,” CVPR, 2019.

[4] F. Schroff et al., ”FaceNet: A Unified Embedding for Face Recognition and Clustering,” CVPR, 2015.

[5] R. Hadsell et al., ”Dimensionality Reduction by Learning an Invariant Mapping,” CVPR, 2006.