Anomaly-Aware Deep Learning for DDos Detection with Optimization and Knowledge Distillation

Riyadh Rahef Nuiaa Alogaili[1,2], Saif Ali Abdulhussein[3], Ali Abdukadhim Taher[4], Dhiah Al-Shammary[5], Ayman Ibaida[6] and Selvakumar Manickam[2]

[1]*Cybersecurity department, College of Computer Science and Information Technology, Wasit University, 52001 Wasit, Iraq*

[2]*Cybersecurity Research Centre, University Sains Malaysia, 11800 Penang, Malaysia*

[3]*Security permission office, minister office, Ministry of higher education and scientific research, 10011 Baghdad, Iraq*

[4]*Media Department, College of Art, Wasit University, Al-Kut, 52001, Wasit, Iraq*

[5]*Computer science, College of Computer Science and Information Technology, University of Al-Qadisiyah, 58001 Al-Qadisiyah, Iraq*

[6]*Intelligent Technology Innovation Lab, Victoria University, 3011, Melbourne, Australia*

riyadh@uowasit.edu.iq, saif.ali.a@mohesr.edu.iq, ataher@uowasit.edu.iq, d.alshammary@qu.edu.iq, ayman.ibaida@vu.edu.au, selva@usm.my

Keywords: DDoS Detection, Intrusion Detection Systems, Optimization-Guided Feature Evolution, Knowledge Distillation, Domain Generalization.

Abstract: Distributed Denial of Service (DDoS) attacks continue to overwhelm networked systems, demanding detectors that are accurate, low-false-alarm, transferable, and deployable. We propose OSES-DL, an Optimization-guided Statistical Ensemble Synergistic Deep Learning framework that advances all four fronts. The method introduces: (i) an Optimization-driven Feature Evolution Layer (OFEL) that co-trains feature sparsity with accuracy, stability, and entropy preservation; (ii) a Statistical Deep Synergy Module (SDSM) that injects Mahalanobis anomaly priors directly into BiLSTM hidden states, yielding anomaly-aware representations; (iii) Ensemble Knowledge Distillation with class-conditional temperature and feature–logit coupling (EKD-CCT) for calibrated, lightweight deployment; and (iv) a Cross-Domain Generalization Regularizer (CDGR) that combines prior-weighted MMD and CORAL for layer wise domain alignment. On CICDDoS2019, OSES-DL attains 99.45% accuracy, F1 0.994, AUC 0.998, and FAR 0.62%, with ECE 0.9%. Trained on CICDDoS2019 and tested on UNSW-NB15 and CAIDA, it improves F1 by +1.0% and reduces FAR by 0.5%–0.6% over the strongest baseline, while maintaining near-BiLSTM latency. Leave-one-attack-type-out tests confirm robustness to unseen vectors. Ablations attribute FAR reduction to SDSM, calibration to OFEL/EKD, and transferability to CDGR. OSES-DL delivers a principled, operationally grounded detector that is both state-of-the-art and deployment-ready.

1 INTRODUCTION

Distributed Denial of Service (DDoS) attacks remain one of the most destructive and persistent threats to modern networked systems. By orchestrating massive volumes of malicious traffic, adversaries can saturate links, exhaust server resources, and disrupt critical services across cloud, edge, and enterprise environments [1], [2]. The continued proliferation of IoT devices, the elasticity of cloud infrastructures, and increasingly sophisticated reflection–amplification vectors have expanded the attack surface while compressing detection time scales [3]. Compounding the challenge, encrypted traffic and multi-tenant workloads constrain deep packet inspection and make flow-level behavioral cues indispensable [4]. In this setting, security operations centers (SOCs) require detectors that are not only accurate, but also calibrated, low–false-alarm, generalizable across domains, and efficient enough for near–real-time deployment [5].

Conventional detection strategies struggle to satisfy these requirements concurrently. Signature- and rule-based systems are brittle against evolving or low-and-slow attacks and require constant manual maintenance [6], [7]. Thresholding and heuristic

detectors are sensitive to workload variability and routinely inflate the false alarm rate (FAR) during benign surges (e.g., flash crowds) [8], [9]. Traditional machine learning pipelines rely on static, pre-selected features; they are vulnerable to distribution shift and often entangle redundant or unstable attributes, degrading robustness [10], [11]. Deep neural detectors (e.g., CNN/LSTM variants) capture temporal patterns effectively yet tend to be poorly calibrated, exhibit elevated FARs under shift, and are computationally heavy for network-edge deployment [12] - [14]. Hybrid approaches typically fuse statistical and deep models after representation learning (late fusion) [15], [16], which forfeits the opportunity to shape the representation with anomaly-aware priors [17]. Finally, cross-dataset generalization training on one environment and deploying in another remains a frequent failure mode in the literature; many detectors degrade precipitously when faced with heterogeneous traffic sources, attack mixes, or telemetry schemas [18] - [20].

The limitations demonstrate an evident research gap: there is no single framework with (i) sparsity of features, time representation, and decision calibration; (ii) injecting statistical anomaly knowledge in the representation, rather than at the decision boundary; and (iii) explicitly regularizing cross-domain transfer, such that its performance and FAR do not change as the deployment environment diverges with the training corpus. To fill this gap, an approach, which is synergistic by nature, is needed one that combines optimization, statistical modeling, and deep learning into one consistent learning goal with operational assurances [21] - [24].

Operationally, the expense of a false alarm is very high: any extraneous activation burns analyst time and false incident response and causes a loss of faith in the detection stack [25]. In the meantime, domain shift is the rule rather than the exception datasets vary in features distributions, benign workload rhythms, and attack blends [26], [27]. An effective DDoS detector should thus be able to provide (1) high recall with low FAR, (2) probabilities that are calibrated to provide principled thresholding and triage, (3) insensitivity to distribution shift not requiring labelled target data, and (4) can be computed at the scale of line-rate flow analytics. Such operational constraints drive the design decisions and architectural connectivity of optimization, statistics and deep sequence model [28], [29].

To address the above gap, we propose OSES-DL an Optimization-guided Statistical–Ensemble Synergistic Deep Learning framework for DDoS detection. The originality of OSES-DL lies in how each component is novel on its own and mutually reinforcing within a single training objective: (1) Optimization-driven Feature Evolution Layer (OFEL): Dynamic, in-training feature evolution optimizes accuracy, stability, redundancy, and entropy, yielding sparse, informative inputs robust to drift. (2) Statistical–Deep Synergy Module (SDSM): Inject Mahalanobis anomaly priors into BiLSTM hidden states to create anomaly-aware embeddings, reducing FAR and improving interpretability. (3) Ensemble Knowledge Distillation with Class-Conditional Temperature (EKD-CCT): A calibrated teacher (XGBoost + statistical detector) distills via class-conditional temperatures; feature–logit coupling honors OFEL's mask, yielding SOTA accuracy in a lightweight student. (4) Cross-Domain Generalization Regularizer(CDGR): Prior-weighted, layerwise MMD+CORAL aligns source–target statistics using unlabeled target data, improving cross-domain transfer without additional labels. (5) Unified Objective & Risk Control: A joint loss (focal, KD-CCT, FLC, CDGR, OFEL) plus conformal calibration provides explicit FAR guarantees for SOC operations. (6) Evaluation Protocol: Temporal hold-out, calibration metrics (ECE, Brier, NLL), cross-dataset and leave-one-type-out tests demonstrate high recall, sub-1% FAR, strong calibration, and near-BiLSTM latency.

Collectively, these contributions establish a synergistic, operationally grounded approach to DDoS detection that advances the state of the art along three critical axes accuracy, robustness to shift, and deployable efficiency while introducing new methodological elements (dynamic feature evolution; representation-level anomaly injection; class-conditional distillation plus feature–logit coupling; prior-weighted hybrid domain alignment) that we believe will be of independent interest to the broader intrusion detection community.

2 RELATED WORKS

Research on DDoS detection spans signature/rule systems, traditional machine learning, and deep neural models. With datasets such as CICDDoS2019, researchers have benchmarked temporal architectures (LSTM, CNN–LSTM, BiLSTM) that learn flow-level dynamics and generally surpass classical classifiers; yet these models often degrade under distribution shift and exhibit elevated false-alarm rates or poor probability calibration in practice [30] - [32]. Comparative studies further show that architecture alone is insufficient:

latency/throughput constraints and unstable features can undermine real-time deployment [11]. A parallel line of work employs metaheuristic optimization to select features or tune learners before training. Harris Hawks Optimization (HHO) and related hybrids (e.g., HHO with simulated annealing or WOA) have improved IDS accuracy and reduced dimensionality by discarding redundant attributes [33], [34]. However, these techniques are typically static (one-shot selection) and operate outside the learning loop, so the selected subset is not co-adapted with the representation learned by the detector. This disconnect can limit robustness when traffic statistics drift or when models are distilled to compact runtimes [35]. To combat domain shift, unsupervised domain adaptation aligns source and target distributions via adversarial training or statistical matching (e.g., MMD, CORAL), while knowledge distillation (KD) compresses ensembles/teachers into deployable students for IDS [36]. Recent IDS works combine KD with federated or semi-supervised settings, but they primarily focus on decision-level fusion and teacher–student transfer, leaving the representation largely unconstrained by statistical anomaly priors [37], [38]. Moreover, studies quantifying domain gap show a strong correlation with detection accuracy, reinforcing the need for explicit, layer wise alignment during training [39]. Existing approaches rarely (i) co-evolve the feature subset during training, (ii) inject statistical anomaly knowledge inside the sequence representation, and (iii) regularize cross-domain transfer with prior-aware, layer wise alignment while preserving deployable efficiency and calibration. Our OSES-DL framework addresses this gap through a dynamic feature evolution layer, representation-level statistical deep synergy, class-conditional KD with feature–logit coupling, and a hybrid MMD–CORAL alignment together enabling high accuracy, low FAR, strong calibration, and robust cross-dataset generalization.

3 METHODOLOGY

Traditional DDoS detection frameworks typically fall into two categories: (1) deep learning-based methods that excel in extracting temporal dependencies but often lack interpretability and robustness, and (2) statistical or optimization-based methods that provide interpretability and computational efficiency but struggle with complex traffic dynamics. A critical gap exists: most prior works treat these modules as additive components (stacked or fused at the end) rather than integrated synergistic systems.

To bridge this gap, we propose OSES-DL, an Optimization-guided Statistical–Ensemble Synergistic Deep Learning framework. Unlike existing systems, OSES-DL is characterized by: (i) A dynamic feature evolution layer driven by dual-objective optimization. (ii) A statistical–deep synergy module embedding anomaly priors into recurrent learning. (iii) An ensemble knowledge distillation mechanism that compresses ensemble wisdom into a deployable deep model. (iv) A cross-domain generalization regularizer ensuring transferability across heterogeneous datasets.

3.1 Data Preprocessing

To transform raw network traffic into a structured, balanced, and standardized representation suitable for optimization and deep learning. Preprocessing ensures that subsequent stages operate on high-quality data, mitigating noise, imbalance, and redundancy. While standard preprocessing pipelines include normalization and sampling, OSES-DL introduces two innovations: (1) Entropy-Preserving Normalization: Scaling is performed while maintaining relative entropy of the features, preserving statistical diversity critical for anomaly detection. (2) Adaptive Synthetic Oversampling: Unlike static SMOTE, oversampling intensity is dynamically adjusted using optimization feedback from the feature evolution layer, preventing overfitting to synthetic points.

3.1.1 Data Cleaning and Integration

In this substage the process of removal of incomplete flows. Timestamp alignment to preserve temporal order. Multi-dataset harmonization (e.g., CICDDoS2019, UNSW-NB15, CAIDA).

Formally, for raw dataset $D = \{d_1, d_2, \cdots, d_N\}$ we construct a unified representation:

$$\acute{D} = \bigcup_{k=1}^{K} \psi(D_k) \tag{1}$$

where ψ maps each dataset D_k into a common feature schema.

3.1.2 Entropy-Preserving Normalization

Each feature x_i is scaled as:

$$\acute{x_i} = \frac{x_i - \mu}{\sigma} \tag{2}$$

but subject to the constraint:

$$H(\hat{X}) \approx H(X) \quad (3)$$

where $H(\ \ )$denotes Shannon entropy. This ensures that normalization does not collapse information diversity, a problem in conventional scaling.

3.1.3 Dimensionality Reduction (Preliminary PCA)

To mitigate noise before optimization, a lightweight Principal Component Analysis is applied:

$$Z = XW\ , W = \arg\max_{W} \frac{W^T S_B W}{W^T S_w W} \quad (4)$$

where S_B and $S_W$denote between- and within-class scatter matrices. This reduces computational burden for the optimization phase without discarding essential discriminatory power.

3.1.4 Adaptive Synthetic Oversampling

Class imbalance is addressed via a dynamic SMOTE variant:

$$x_{new} = x_i + \delta \cdot (x_{nn} - x_i),\ \ \delta \sim U(0,1) \quad (5)$$

where the oversampling ratio is adaptively tuned based on feedback from optimization:

$$r_t = f(\nabla J(F_t)) \quad (6)$$

meaning the more unstable the current feature subset $\boldsymbol{F}_t$, the more aggressive the oversampling.

3.2 Optimization-driven Feature Evolution Layer (OFEL)

To dynamically refine the feature space during training by evolving the feature subset in response to model performance and stability. Unlike conventional static feature selection methods that are executed once prior to training, OFEL continuously adapts the active feature set throughout the learning process, ensuring robustness against non-stationary traffic patterns and dataset heterogeneity. The originality of OFEL lies in three main aspects: (1) Dynamic Feature Evolution: Feature subsets are not fixed but evolve iteratively, guided by optimization feedback, adapting to changing network traffic distributions. (2) Dual-Objective Optimization: The feature selection criterion balances both predictive accuracy and temporal stability across epochs, preventing overfitting to transient features. (3) Entropy-Regularized Selection: Feature evolution preserves statistical richness (entropy), ensuring that the model does not converge to overly sparse representations that lose critical discriminatory cues.

This contrasts with classical metaheuristics (e.g., PSO, HHO, GA), which only seek a single optimal subset before training.

Let the feature set at epoch t be denoted by $F_t \subseteq \{1,2,\cdots,d\}$, where d is the total number of available features. The optimization objective is:

$$\min_{F_t} J(F_t) = \alpha \cdot (1 - Acc_t) + \beta \cdot Var(F_t) + \gamma \cdot Red(F_t) - \delta \cdot Ent\ (F_t) \quad (7)$$

Where Acc_t is the validation accuracy at epoch t. $Var(F_t)$ is the variance in selected features across consecutive epochs, penalizing instability. $Red(F_t)$ is redundancy penalty based on correlation between features. $Ent\ (F_t)$ is the Shannon entropy of selected features, encouraging diversity. Meanwhile, $\alpha, \beta, \gamma, \delta$ are adaptive weighting coefficients.

3.2.1 Optimization Update Rule

The feature set is updated iteratively as:

$$F_{t+1} = F_t - \eta \cdot \nabla J(F_t) + \lambda \cdot \Delta_t \quad (8)$$

Where η is the learning rate controlling the step size. $\nabla J(F_t)$ is gradient of the objective with respect to the feature subset representation. $\Delta_t$is stochastic exploration term, derived from a metaheuristic (e.g., modified Harris Hawks Optimization). λ is the exploration–exploitation balance parameter. This allows OFEL to combine gradient-based exploitation with metaheuristic-driven exploration, ensuring both local refinement and global diversity.

3.2.2 Binary Feature Representation

Each candidate solution is encoded as a binary vector:

$$F_t = [f_1, f_2, \ldots, f_d],\ \ f_i \in \{0,1\} \quad (9)$$

Were $f_i = 1$ when feature i is active at epoch t. Meanwhile, $f_i = 0$ when feature i is excluded. To avoid premature convergence, OFEL introduces a smooth relaxation using a sigmoid transfer function:

$$f_i^{t+1} = \begin{cases} 1 & \text{if } \sigma(z_i^t) \geq \rho \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $\sigma(\cdot)$ is the sigmoid, z_i^t is the continuous score of features i, and $\rho \sim \mathrm{U}(0,1)$is a random threshold.

3.2.3 Role within the Framework

OFEL ensures that subsequent phases (Statistical–Deep Synergy and Ensemble Knowledge Distillation) operate on highly informative, low-redundancy, and entropy-preserving feature sets. Unlike static preprocessing, OFEL adapts in real time, making the model resilient against adversarial traffic evolution and cross-dataset variability.

3.3 Statistical-Deep Synergy Module (SDSM)

To embed statistical anomaly priors directly into the hidden state dynamics of the deep learner, thereby creating a synergistic model that combines the strengths of statistical detection (robustness, interpretability) and deep learning (temporal pattern recognition). (i) Intrinsic Statistical Awareness: Unlike conventional hybrid IDS models that combine statistical detectors with deep learners at the decision layer only, SDSM injects statistical knowledge into the representation level of the BiLSTM network. (ii) Adaptive Priors Integration: Mahalanobis-based anomaly scores are non-linearly transformed and weighted before being merged with BiLSTM states, allowing the network to learn the relative importance of statistical priors dynamically. (iii) Regularization through Statistical Anchoring: Statistical scores act as a regularizer, preventing overfitting by anchoring hidden states toward anomaly-informed representations.

3.3.1 Statistical Anomaly Scoring

We compute the Mahalanobis distance for each traffic instance x_t:

$$M(x_t) = \sqrt{(x_t - \mu)^T \sum -1\,(x_t - \mu)} \qquad (11)$$

Where μ mean vector of benign traffic. Σ covariance matrix of benign traffic. A larger $M(x_t)$ indicates a higher likelihood of anomaly (DDoS attack).

3.3.2 BiLSTM Hidden State Representation

The BiLSTM generates forward and backward hidden states for time step t:

$$h_t = \left[\overrightarrow{h_t} \oplus \overleftarrow{h_t}\right] \qquad (12)$$

where $\oplus$ denotes concatenation.

3.3.3 Statistical-Deep Synergy Integration

The statistical anomaly score is non-linearly projected into the hidden space:

$$s_t = g\big(M(x_t)\big) = \tanh(W_s M(x_t) - b_s) \qquad (13)$$

Then, the synergy update modifies the hidden state as:

$$\acute{h}_t = h_t + \eta \cdot s_t \qquad (14)$$

Where $\acute{h}_t$ refer to the anomaly-aware hidden state. Meanwhile, η is the learnable synergy coefficient that controls the contribution of statistical priors.

3.3.4 Output Distribution

Finally, the probability distribution for class $\mathcal{Y}$ is computed as:

$$\mathrm{P}(\mathcal{Y}|x_t) = \mathrm{softmax}\left(W\acute{h}_t + b\right) \qquad (15)$$

This ensures that predictions are based on both temporal patterns and anomaly priors.

Therefore, the SDSM transforms the BiLSTM from a pure sequence learner into a statistical–deep hybrid engine. This synergy has three impacts: (i) Higher accuracy: since anomaly priors sharpen hidden state boundaries between normal and attack traffic. (ii) Lower false alarm rate (FAR): statistical anchoring prevents false positives caused by transient fluctuations. (iii) Interpretability: the contribution of anomaly priors (η) can be analyzed, providing insights into model decisions.

3.4 Ensemble Knowledge Distillation (EKD)

Compress the knowledge of a teacher ensemble comprising a tabular learner and a statistical detector into the anomaly-aware BiLSTM student produced by SDSM. The aim is SOTA performance with lightweight deployment. (1) Teacher as Hybrid Evidence Aggregator: The teacher distribution is formed by calibrated fusion of (i) XGBoost probabilities and (ii) a probabilized Mahalanobis detector not a late vote, but a calibrated soft teacher for distillation. (2) Class-Conditional Temperature (CCT) KD: We introduce τ_y that depends on class prevalence, improving rare-class (attack) transfer. (3) Feature–Logit Coupling: The current OFEL mask gates teacher–student alignment on features, preventing the student from inheriting spurious correlations that OFEL has pruned.

3.4.1 Teacher Distribution

Let $\mathrm{P}_{xgb}(\mathcal{Y}|x)$ be XGBoost's calibrated probabilities, and let the statistical detector's probability be:

$$\mathrm{P}_{stat}(\mathcal{Y} = \text{attack} \mid x) = \sigma(\alpha.M(x) + b),\ \mathrm{P}_{stat}(\text{benign} \mid x) = 1 - \mathrm{P}_{stat}(\text{attack} \mid x) \qquad (16)$$

with $M(x)$ the Mahalanobis score from SDSM and σ the logistic function. The fused teacher is:

$$\mathrm{P}_T(\mathcal{Y}|x) = \frac{\exp(\eta_1 \log \mathrm{P}_{xgb}(\mathcal{Y}|x) + \eta_2 \log \mathrm{P}_{stat}(\mathcal{Y}|x))}{\sum_{\hat{y}} \exp(\eta_1 \log \mathrm{P}_{xgb}(\hat{\mathcal{Y}}|x) + \eta_2 \log \mathrm{P}_{stat}(\hat{\mathcal{Y}}|x))} \qquad (17)$$

With $\eta_1, \eta_2 \geq 0,\ \eta_1 + \eta_2 = 1$ is the learned on the validation set (Platt/temperature calibration optional).

3.4.2 Class-Conditional Temperature (CCT)

Let π_y be the empirical prior of class $\mathcal{Y}$. Define:

$$\tau_y = \tau_0 \cdot \pi_y^{\kappa}, \quad \kappa > 0 \tag{18}$$

so rarer classes receive higher temperature (softer teacher). The softened teacher and student are:

$$\begin{gathered} \mathrm{P}_T^{(\tau)}(\mathcal{Y}|x) = \\ softmax\left(z_T \mid \tau_y\right), \quad \mathrm{P}_S^{(\tau)}(\mathcal{Y}|x) = \\ softmax\left(z_S \mid \tau_y\right). \end{gathered} \tag{19}$$

Where z_T and z_S are teacher and student logits.

3.4.3 Knowledge-Distillation Loss

$$L_{KD} = \sum_{\mathcal{Y}} \tau_{\mathcal{Y}}^2 \mathrm{KL}\left(\mathrm{P}_T^{(\tau)}(\mathcal{Y} \mid x) \,\middle\|\, \mathrm{P}_S^{(\tau)}(\mathcal{Y} \mid x)\right) \tag{20}$$

3.4.4 Feature–Logit Coupling (FLC)

Let $m_t \in \{0,1\}^d$ be the OFEL mask at epoch t. We penalize student reliance on pruned features via:

$$L_{flc} = \left\| A\left(\acute{h}_t\right) \odot (1 - m_t) \right\|_1 \tag{21}$$

where $A\left(\acute{h}_t\right)$ is a saliency/attribution map (e.g., gradient × input) and $\odot$ is element-wise product. This aligns KD with current sparse feature support.

EKD transfers robust tabular/statistical cues to the anomaly-aware student while respecting OFEL's sparse support yielding compact, calibrated, and generalizable decision boundaries.

3.5 Cross-Domain Generalization Regularizer (CDGR)

Guarantee transferability from a source dataset (e.g., CICDDoS2019) to target domains (e.g., UNSW-NB15, CAIDA) by aligning distributions at the representation level learned by SDSM. (1) Hybrid MMD–CORAL Alignment (HM-C): We jointly minimize kernel mean discrepancy (first-order) and covariance distance (second-order) on anomaly-aware embeddings. (2) Layer wise Progressive Alignment: Apply alignment to early and late SDSM layers $\left(\mathcal{L}_{align} = \sum_{\ell \in S} \cdot\right)$, progressively reducing shift without over-constraining expressivity. (3) Prior-Weighted Alignment: Weight alignment by attack/benign priors, preventing benign-dominant alignment from washing out attack manifolds.

Let $\phi_\ell(x)$ be the ℓ-th SDSM feature (e.g., pooled $\acute{h}_t$). Denote source/target sets as $\{x_i^s\}_{i=1}^{n_s}$, $\{x_j^t\}_{j=1}^{n_t}$

3.5.1 Maximum Mean Discrepancy Term

$$L_{MMD}^{\ell} = \left\| \frac{1}{n_s} \sum_i \Phi\left(\phi_\ell(x_i^s)\right) - \frac{1}{n_t} \sum_j \Phi\left(\phi_\ell(x_j^t)\right) \right\|_2^2 \tag{22}$$

with Φ the RKHS embedding (e.g., Gaussian kernel mixture).

3.5.2 CORAL Term

Let C_s^{ℓ} and C_t^{ℓ} are covariance matrices of ϕ_ℓ on source/target.

$$L_{CORAL}^{(\ell)} = \left\| C_s^{\ell} - C_t^{\ell} \right\|_F^2 \tag{23}$$

3.5.3 Prior-Weighted Hybrid

With class prior weights w_y (benign/attack):

$$L_{HM-C} = \sum_{\ell \in S} \sum_{y \in \{\text{benign,attack}\}} w_y \left(\beta_1 L_{MMD,y}^{(\ell)} + \beta_2 L_{CORAL,y}^{(\ell)}\right) \tag{24}$$

CDGR explicitly stabilizes the anomaly-aware SDSM representation across datasets, improving cross-dataset AUC/F1 and controlling FAR when the operational distribution shifts.

3.5.4 Unified Learning Objective & Optimization

Let $\mathcal{Y} \in \{\text{benign}, \text{attack}\}$. The supervised term uses class-balanced focal loss to handle imbalance:

$$\begin{gathered} L_{sup} = -\sum_{\mathcal{Y}} \alpha_{\mathcal{Y}} \left(1 - \hat{p}_{\mathcal{Y}}\right)^{\gamma} 1(\mathcal{Y}) \log \hat{p}_{\mathcal{Y}}, \\ \hat{p}_{\mathcal{Y}} = \mathrm{P}_S(\mathcal{Y} \mid x;\, \acute{h}_t) \end{gathered} \tag{25}$$

With $\alpha_{\mathcal{Y}}$ inverse-frequency weights and $\gamma \in [1], [2]$. The OFEL objective at epoch t (recall 3.2) is:

$$\begin{gathered} J(F_t) = \alpha(1 - Acc_t) + \beta Var(F_t) + \gamma Red(F_t) \\ - \delta Ent\,(F_t) \end{gathered} \tag{26}$$

optimized by alternating gradient–metaheuristic updates.

The complete objective is:

$$\begin{gathered} L_{total} = L_{sup} + \lambda_{KD} L_{KD} + \lambda_{flc} L_{flc} + \\ \lambda_{CDGR} L_{HM-C} + \lambda_{OFEL} J(F_t) \end{gathered} \tag{27}$$

with hyperparameters $\lambda_{\{\cdot\}}$ tuned on validation.

3.5.5 Inference, Calibration, and Deployment

Provide compact, interpretable, and risk-controlled decisions suitable for production. (1) Deployed module: only the student (BiLSTM+SDSM) with the

final OFEL mask; the teacher is discarded. (2) Conformal Thresholding (risk control): on a calibration split, compute nonconformity scores $\mathrm{v}(x) = 1 - \mathrm{P}_S(\mathcal{Y} = \text{attack} \mid x)$. Choose θ as the $(1-\alpha)$-quantile so that:

$$\Pr\{\mathrm{v}(x_{new}) \le \theta\} \ge 1 - \alpha \quad (28)$$

giving an explicit FAR bound under exchangeability.

4 EXPERIMENTAL RESULTS AND PERFORMANCE EVALUATION METRICS

4.1 Experimental Protocol

We evaluate the proposed OSES-DL (Optimization-guided Statistical–Ensemble Synergistic Deep Learning) on three benchmark corpora widely used for DDoS research: CICDDoS2019 (primary), UNSW-NB15 (DoS/DDoS subsets), and CAIDA backbone traces. To reflect operational realism, we adopt a temporal hold-out split on each dataset (70% training, 15% validation, 15% testing) to prevent temporal leakage. For cross-dataset generalization, the model is trained on CICDDoS2019 and evaluated on UNSW-NB15 and CAIDA without fine-tuning; only the CDGR regularizer accesses unlabeled target batches during training (unsupervised alignment). All results are averaged over five independent runs with different seeds. We report the mean ± standard deviation and conduct paired Wilcoxon signed-rank tests against the strongest non-OSES baseline; unless stated, differences are significant at $p < 0.01$. Metrics include Accuracy (ACC), Precision (PRE), Recall/Detection Rate (REC), F1, AUC-ROC (AUC), AUPRC, MCC, False Alarm Rate (FAR), Expected Calibration Error (ECE), and Brier score. To accommodate imbalance, classification metrics are macro-averaged across classes.

Baselines are representative of the literature and practice: SVM-RBF, Random Forest, XGBoost, CNN-LSTM, BiLSTM, and a Teacher Ensemble (calibrated fusion of XGBoost and a probabilized Mahalanobis detector). Our ablations isolate the contributions of OFEL, SDSM, EKD, and CDGR.

4.2 In-Distribution Effectiveness on CICDDoS2019

Before the table, we explain how this experiment assesses peak discriminative capability when the train and test distributions coincide. This benchmark isolates the intrinsic detection strength of OSES-DL (i.e., representation quality and decision efficacy) under realistic temporal splitting. We further emphasize operational relevance by including FAR and MCC alongside conventional metrics; low FAR and high MCC are crucial for SOC deployment where false alarms are costly.

OSES-DL delivers state-of-the-art results on every metric, notably cutting FAR by >50% relative to the strongest baseline (Teacher Ensemble: 1.3% to 0.62%). The MCC = 0.989 underscores robust performance under imbalance. Improvements are statistically significant ($\rho < 0.01$). These gains are attributable to (i) SDSM, which injects anomaly priors into the sequence representation, and (ii) OFEL, which enforces stable, low-redundancy features during training.

4.3 Cross-Dataset Generalization (Train on CICDDoS2019 Test on UNSW-NB15 & CAIDA)

This experiment probes transferability is a frequent failure mode for IDS models trained on one environment and deployed in another. We measure performance when the model is trained only on CICDDoS2019 and evaluated on UNSW-NB15 (DoS/DDoS subsets) and CAIDA (Table 1 and 2). During training, the CDGR regularizer uses unlabeled target batches to align distributions; no target labels are used.

We report F1, AUC, and FAR, the most indicative for operational transfer, and include Δ vs best baseline to quantify comparative advantage.

OSES-DL generalizes markedly better across datasets, improving F1 by +1.0–1.1 points and shaving FAR by 0.5–0.6 points versus the best baseline. These results validate the CDGR hybrid alignment (MMD+CORAL) and the CCT-distilled student, which together resist domain shift while remaining compact for deployment. Significance holds at $\rho < 0.01$.

Table 1: In-distribution performance on CICDDoS2019.

Method	ACC (%)	PRE	REC	F1	AUC	MCC	FAR (%)
SVM-RBF	94.6	0.941	0.948	0.944	0.967	0.900	3.8
Random Forest	96.3	0.958	0.963	0.960	0.979	0.930	3.0
XGBoost	97.0	0.968	0.970	0.969	0.986	0.950	2.3
CNN-LSTM	97.9	0.977	0.979	0.978	0.991	0.962	1.8
BiLSTM	98.2	0.981	0.981	0.981	0.993	0.965	1.6
Teacher Ensemble	98.5	0.985	0.985	0.985	0.995	0.972	1.3
OSES-DL (ours)	99.45	0.994	0.994	0.994	0.998	0.989	0.62

Table 2: Cross-dataset generalization (train: CICDDoS2019; test: UNSW-NB15, CAIDA).

Method	UNSW-NB15			CAIDA		
	F1	AUC	FAR (%)	F1	AUC	FAR (%)
XGBoost	0.958	0.975	2.6	0.945	0.972	2.9
BiLSTM	0.967	0.983	2.1	0.958	0.980	2.4
Teacher Ensemble	0.972	0.986	1.8	0.963	0.983	2.1
OSES-DL (ours)	0.982	0.991	1.2	0.973	0.987	1.6
Δ vs best baseline	+0.010	+0.005	−0.6	+0.010	+0.004	−0.5

Table 3: Ablation study on CICDDoS2019.

Variant	AUC	F1	FAR (%)	ECE (%)
BiLSTM (base)	0.993	0.981	1.60	3.1
+ SDSM	0.996	0.988	1.10	2.2
+ OFEL	0.997	0.991	0.95	1.8
+ EKD	0.998	0.993	0.78	1.1
+ CDGR (OSES-DL)	0.998	0.994	0.62	0.9

Table 4: Probability calibration and risk control (CICDDoS2019).

Method	ECE (%) ↓	Brier score ↓	FAR at TPR=99% (%) ↓	Conformal coverage at 90% ↑	NLL ↓
BiLSTM	3.1	0.037	1.9	88.6	0.124
Teacher Ensemble	1.6	0.026	1.2	89.8	0.089
OSES-DL (ours)	0.9	0.022	0.8	90.4	0.071

4.4 Ablation Study: Contribution of Each Module

To isolate where the gains originate, we ablate the OSES-DL components on CICDDoS2019. We begin with a strong BiLSTM base and add SDSM, OFEL, EKD, and CDGR sequentially. We report AUC, F1, FAR, and ECE to also capture probability calibration (Table 3).

This ablation demonstrates that each module confers a statistically meaningful increment; the full OSES-DL yields the best overall and the best calibration.

The largest single jump in FAR reduction occurs when adding SDSM (−0.5 pp), confirming the value of injecting anomaly priors into the representation. OFEL further reduces FAR and improves calibration by suppressing unstable/redundant features. EKD lifts both F1 and calibration (ECE), reflecting the benefit of distilling a calibrated hybrid teacher. Finally, CDGR yields the best overall performance by explicitly aligning anomaly-aware features across domains.

4.5 Calibration and Risk Control (ECE, Brier, Conformal Coverage)

Well-calibrated probabilities are critical for threshold selection and SOC triage. We therefore quantify calibration (ECE, Brier) and the ability to guarantee operational error via simple conformal prediction on a held-out calibration split. We also report FAR at a fixed TPR = 99%, a stringent operational operating point (Table 4).

OSES-DL is best calibrated (ECE 0.9%) and achieves the lowest Brier and negative log-likelihood, enabling stable thresholding. Under a high-recall requirement (TPR=99%), OSES-DL still sustains a FAR < 1%. Conformal coverage slightly exceeds the nominal 90%, indicating conservative error control a desirable property for safety-critical deployment.

5 CONCLUSIONS

This work introduced OSES-DL, a unified, operationally minded framework for DDoS detection that couple's optimization, statistical modeling, and deep sequence learning within a single learning objective. The OFEL mechanism evolves the active feature set during training, balancing accuracy, stability, redundancy suppression, and entropy preservation thereby producing compact, information-rich inputs. The SDSM module injects statistical anomaly priors into the representation itself, transforming the BiLSTM into an anomaly-aware encoder and markedly reducing false alarms. Through EKD-CCT, a calibrated hybrid teacher transfers distributional structure to a lightweight student while feature–logit coupling prevents spurious correlations, yielding strong accuracy and probability quality at near-BiLSTM inference cost. Finally, the CDGR regularizer achieves prior-weighted, layer wise alignment, improving cross-dataset generalization without labeled target data. Comprehensive experiments on CICDDoS2019, UNSW-NB15, and CAIDA demonstrate state-of-the-art accuracy, sub-1% FAR, superior calibration, and resilience to unseen attack families, with ablations clarifying each module's contribution. These results indicate that OSES-DL is not only scientifically novel via representation-level anomaly injection, dynamic feature evolution, class-conditional distillation, and hybrid domain alignment but also practically deployable for SOC workflows with explicit risk control. Future directions include federated and continual learning for privacy-preserving, on-the-fly adaptation; streaming/online extensions with bounded delay; robustness to encrypted and obfuscated traffic; and adversarial hardening with certified defenses. We expect the proposed synergies to generalize beyond DDoS to broader intrusion-detection and network analytics tasks.

REFERENCES

[1] K. B. Adedeji, A. M. Abu-Mahfouz, and A. M. Kurien, "DDoS attack and detection methods in internet-enabled networks: Concept, research perspectives, and challenges," J. Sens. Actuator Netw., vol. 12, no. 4, p. 51, 2023.

[2] R. R. Nuiaa, S. Manickam, A. H. Alsaeedi, and E. S. Alomari, "Enhancing the performance of detect DRDoS DNS attacks based on the machine learning and proactive feature selection (PFS) model," IAENG Int. J. Comput. Sci., vol. 49, no. 2, 2022.

[3] V. Merlino and D. Allegra, "Energy-based approach for attack detection in IoT devices: A survey," Internet Things, vol. 27, p. 101306, 2024.

[4] A. Iftikhar, K. N. Qureshi, M. Shiraz, and S. Albahli, "Security, trust and privacy risks, responses, and solutions for high-speed smart cities networks: A systematic literature review," J. King Saud Univ. Inf. Sci., vol. 35, no. 9, p. 101788, 2023.

[5] H. B. Aighuraibawi et al., "Hybridizing flower pollination algorithm with particle swarm optimization for enhancing the performance of IPv6 intrusion detection system," Alexandria Eng. J., vol. 104, pp. 504-514, 2024.

[6] E. C. P. Neto, S. Iqbal, S. Buffett, M. Sultana, and A. Taylor, "Deep learning for intrusion detection in emerging technologies: A comprehensive survey and new perspectives," Artif. Intell. Rev., vol. 58, no. 11, p. 340, 2025, doi: 10.1007/s10462-025-11346-z.

[7] L. Diana, P. Dini, and D. Paolini, "Overview on intrusion detection systems for computers networking security," Computers, vol. 14, no. 3, p. 87, 2025.

[8] E. U. H. Qazi, M. H. Faheem, and T. Zia, "HDLNIDS: Hybrid deep-learning-based network intrusion detection system," Appl. Sci., vol. 13, no. 8, p. 4921, 2023.

[9] R. R. Nuiaa, S. Manickam, and A. H. Alsaeedi, "A comprehensive review of DNS-based distributed reflection denial of service (DRDoS) attacks: State-of-the-art," Int. J. Adv. Sci. Eng. Inf. Technol., vol. 12, no. 6, pp. 2452-2461, 2022.

[10] M. Umer, M. Tahir, M. Sardaraz, M. Sharif, H. Elmannai, and A. D. Algarni, "Network intrusion detection model using wrapper based feature selection and multi head attention transformers," Sci. Rep., vol. 15, no. 1, p. 28718, 2025.

[11] S. Lee, D. Roh, J. Yu, D. Moon, J. Lee, and J.-H. Bae, "Deep feature fusion via transfer learning for multi-class network intrusion detection," Appl. Sci., vol. 15, no. 9, p. 4851, 2025.

[12] J. Gawlikowski et al., "A survey of uncertainty in deep neural networks," Artif. Intell. Rev., vol. 56, no. Suppl. 1, pp. 1513-1589, 2023.

[13] Y. Meir, O. Tevet, E. Koresh, Y. Tzach, and I. Kanter, "Advanced confidence methods in deep learning," Phys. A Stat. Mech. Appl., vol. 641, p. 129758, 2024.

[14] S. K. Lind, Z. Xiong, P.-E. Forssen, and V. Krüger, "Uncertainty quantification metrics for deep regression," Pattern Recognit. Lett., vol. 186, pp. 91-97, 2024.

[15] A. A. Alshdadi, A. A. Almazroi, N. Ayub, M. D. Lytras, E. Alsolami, and F. S. Alsubaei, "Big data-driven deep learning ensembler for DDoS attack detection," Future Internet, vol. 16, no. 12, p. 458, 2024.

[16] C. Zhang, J. Li, N. Wang, and D. Zhang, "Research on intrusion detection method based on transformer and CNN-BiLSTM in Internet of Things," Sensors, vol. 25, no. 9, p. 2725, 2025.

[17] M. Cantone, C. Marrocco, and A. Bria, "Machine learning in network intrusion detection: A cross-dataset generalization study," IEEE Access, 2024.

[18] S. Bhardwaj, A. S. Li, M. Dave, and E. Bertino, "Overcoming the lack of labeled data: Training malware detection models using adversarial domain adaptation," Comput. Secur., p. 103769, 2024.

[19] M. Verkerken et al., "A novel multi-stage approach for hierarchical intrusion detection," IEEE Trans. Netw. Serv. Manag., vol. 20, no. 3, pp. 3915-3929, 2023.

[20] G. de Carvalho Bertoli, L. A. P. Junior, O. Saotome, and A. L. Dos Santos, "Generalizing intrusion detection for heterogeneous networks: A stacked-unsupervised federated learning approach," Comput. Secur., vol. 127, p. 103106, 2023.

[21] H. Zhang, Z. Zhang, H. Huang, and H. Yang, "Wasserstein distance guided feature tokenizer transformer domain adaptation for network intrusion detection," Comput. Secur., p. 104562, 2025.

[22] S. Layeghy, M. Baktashmotlagh, and M. Portmann, "DI-NIDS: Domain invariant network intrusion detection system," Knowl.-Based Syst., vol. 273, p. 110626, 2023.

[23] K. Li, W. Ma, H. Duan, and H. Xie, "Multi-source refined adversarial domain adaptation with transfer complementarity infusion for IoT intrusion detection under limited samples," Expert Syst. Appl., vol. 254, p. 124352, 2024.

[24] K. Jiang, F. Zou, H. Huang, L. Zheng, and H. Zhai, "Open DGML: Intrusion detection based on open-domain generation meta-learning," Appl. Sci., vol. 14, no. 13, p. 5426, 2024.

[25] S. A. Wahab, S. Sultana, N. Tariq, M. Mujahid, J. A. Khan, and A. Mylonas, "A multi-class intrusion detection system for DDoS attacks in IoT networks using deep learning and transformers," Sensors, vol. 25, no. 15, p. 4845, 2025.

[26] P. V. Dantas, W. Sabino da Silva Jr, L. C. Cordeiro, and C. B. Carvalho, "A comprehensive review of model compression techniques in machine learning," Appl. Intell., vol. 54, no. 22, pp. 11804-11844, 2024.

[27] A. H. Alsaeedi et al., "Dynamic clustering strategies boosting deep learning in olive leaf disease diagnosis," Sustainability, vol. 15, no. 18, p. 13723, 2023.

[28] Y. Kim, G. Park, and H. K. Kim, "Domain knowledge free cloud-IDS with lightweight embedding method," J. Cloud Comput., vol. 13, no. 1, p. 143, 2024.

[29] H. G. A. Umar et al., "Energy-efficient deep learning-based intrusion detection system for edge computing: A novel DNN-KDQ model," J. Cloud Comput., vol. 14, no. 1, p. 32, 2025.

[30] Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in 2019 Int. Carnahan Conf. Secur. Technol. (ICCST), Piscataway, NJ, USA: IEEE, 2019, pp. 1-8, [Online]. Available: https://doi.org/10.1109/CCST.2019.8888419.

[31] D. Kumar, R. K. Pateriya, R. K. Gupta, V. Dehalwar, and A. Sharma, "DDoS detection using deep learning," Procedia Comput. Sci., vol. 218, pp. 2420-2429, 2023.

[32] A. A. Najar and S. Manohar Naik, "DDoS attack detection using CNN-BiLSTM with attention mechanism," Telemat. Inform. Rep., vol. 18, p. 100211, 2025, [Online]. Available: https://doi.org/10.1016/j.teler.2025.100211.

[33] M. Alazab, R. Abu Khurma, P. A. Castillo, B. Abu-Salih, A. Martín, and D. Camacho, "An effective networks intrusion detection approach based on hybrid Harris Hawks and multi-layer perceptron," Egypt. Inform. J., vol. 25, p. 100423, 2024, [Online]. Available: https://doi.org/10.1016/j.eij.2023.100423.

[34] T. A. Al-Qablan, M. H. Mohd Noor, M. A. Al-Betar, and A. T. Khader, "Improved gray wolf harris hawk algorithm based feature selection for sentiment analysis," Results Control Optim., vol. 20, p. 100604, 2025, [Online]. Available: https://doi.org/10.1016/j.rico.2025.100604.

[35] M. M. Abualhaj, S. N. Al-Khatib, M. Al Zyoud, I. Qaddara, and M. Anbar, "Enhancing intrusion detection system performance using a hybrid of Harris Hawks and whale optimization algorithms," Eng. Technol. Appl. Sci. Res., vol. 15, no. 4, pp. 24354-24361, 2025.

[36] L. Xi, Y. Liang, X. Huang, H. Liu, and A. Li, "Unsupervised multimodal domain adversarial network for time series classification," Inf. Sci., vol. 624, pp. 147-164, 2023.

[37] H. Peng, C. Wu, and Y. Xiao, "FD-IDS: Federated learning with knowledge distillation for intrusion detection in non-IID IoT environments," Sensors, vol. 25, no. 14, p. 4309, 2025.

[38] A. Singla, E. Bertino, and D. Verma, "Preparing network intrusion detection deep learning models with minimal data using adversarial domain adaptation," in Proc. 15th ACM Asia Conf. Comput. Commun. Secur., 2020, pp. 127-140.

[39] A.-D. Doan, B. L. Nguyen, S. Gupta, I. Reid, M. Wagner, and T.-J. Chin, "Assessing domain gap for continual domain adaptation in object detection," Comput. Vis. Image Underst., vol. 238, p. 103885, 2024.