Hierarchical Attention-Based Convolutional Neural Network Model for Intrusion Detection

Rodolfo Martinez Cadena, José Adán Hernández-Nolasco, and Noel Zacarias-Morales

***Abstract*—The increasing scale and complexity of internet-connected systems demand robust intrusion detection under realistic traffic conditions. This study presents H.A.L.C.CO.N (Hierarchical Attention-based Loss Equalization with CatBoost-enhanced Convolutional Neural Network), a multiclass intrusion detection model evaluated on the real-world LITNET-2020 dataset. The model integrates convolutional feature extraction, hierarchical attention, CatBoost-based encoding for high-cardinality categorical features, and Equalization Loss V2 (EQLv2) to address severe class imbalance. Experimental results show strong performance, achieving a detection rate of 99.997%, an F1-score of 99.997%, an accuracy of 99.996%, and a false positive rate of 0.0135%. These findings indicate that H.A.L.C.CO.N is an effective and practical solution for real-world multiclass intrusion detection.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10406

***Index Terms*—Intrusion detection, convolutional neural networks, real data, cybersecurity, deep learning.**

I. Introduction

THE growing interconnectivity of devices and users has expanded the attack surface of modern networks and intensified cybersecurity risks. In this scenario, Intrusion Detection Systems (IDS) are essential for identifying malicious activity and supporting network defense. However, many IDS models are still evaluated on synthetic datasets that fail to capture the variability, noise, and class imbalance of real-world traffic.

Conventional machine learning methods, such as k-NN and SVM [1], [2], as well as deep learning approaches [3], [4], have shown promising results. Still, high false positive rates (FPR) and limited realism in evaluation data remain major barriers to practical deployment. More recent models, such as CANET [5], introduced attention mechanisms and Equalization Loss v2 (EQLv2) [6] to improve learning under class imbalance. However, they were mainly validated on benchmark datasets such as NSL-KDD and UNSW-NB15 [7], which reduces their relevance for operational traffic conditions.

This work presents **H.A.L.C.CO.N**, a multiclass IDS model that combines convolutional learning, hierarchical attention, CatBoost-based encoding, and EQLv2, and evaluates it on the real-world LITNET-2020 dataset [8]. The aim is to improve multiclass intrusion detection under realistic traffic conditions while maintaining low false alarm rates. By focusing on real flow-based traffic, severe class imbalance, and heterogeneous categorical features, H.A.L.C.CO.N addresses key limitations of previous IDS studies.

The associate editor coordinating the review of this manuscript and approving it for publication was Martín Pedemonte (*Corresponding author: Rodolfo Martínez Cadena*).

Rodolfo Martínez Cadena, J. A. Nolasco and N. Zacarias are with Division Académica de Ciencias y Tecnologías de la Información, Universidad Juárez Autónoma de Tabasco, Cunduacán, Mexico (e-mails: 232H21006@alumno.ujat.mx, adan.hernandez@ujat.mx, and noel.zacarias@ujat.mx).

II. Related Work

Intrusion Detection Systems (IDS) have evolved from rule-based approaches to machine learning and deep learning models capable of capturing complex traffic patterns [3], [9]. Conventional methods such as k-nearest neighbors, support vector machines, and ensemble models improved intrusion detection, but often depend on handcrafted features and struggle with the nonlinear structure of modern network traffic [1], [10], [11]. More recent deep learning models, including convolutional, recurrent, and hybrid architectures, have shown stronger representational capacity [4], [12]–[14]. Attention-based models further improved feature selectivity, with CANET being a representative example [5], [15], [16].

A major limitation of IDS research is the reliance on benchmark datasets that do not fully reflect operational traffic. Although NSL-KDD and UNSW-NB15 are useful for comparison, they do not completely capture the variability, noise, heterogeneity, and class imbalance of real networks [7], [9], [17], [18]. By contrast, LITNET-2020 contains real flow-based traffic from an academic network and provides a more demanding benchmark for multiclass IDS evaluation [8].

In this setting, two challenges are particularly relevant: severe class imbalance and high-cardinality categorical features. EQLv2 helps mitigate gradient domination in long-tailed classification [6], while CatBoost-based encoding provides a stronger alternative to basic label or one-hot encoding for categorical attributes [19]. Motivated by these challenges, this study evaluates H.A.L.C.CO.N, a multiclass IDS that combines convolutional learning, hierarchical attention, CatBoost-based encoding, and EQLv2 on LITNET-2020.

III. Methodology

H.A.L.C.CO.N was evaluated as a multiclass intrusion detection model on the real-world **LITNET-2020** dataset. The methodology comprises four main stages: feature selection and preprocessing, categorical encoding, hierarchical attention-enhanced convolutional learning, and imbalance-aware optimization.

A. Dataset and Feature Preparation

LITNET-2020 is a real flow-based network traffic dataset collected in an academic environment and provides a more realistic benchmark than synthetic datasets because it includes heterogeneous traffic patterns, severe class imbalance, and high-cardinality categorical attributes [8]. It contains **13 classes**, including the benign `Normal` class and **12 attack categories**. The `Normal` class accounts for approximately **92%** of the samples, making minority-class detection particularly challenging. Fig. 1 shows the class distribution of LITNET-2020.

Although the complete dataset was explored to verify its global class structure, training and evaluation were performed on a **10%** subset constructed to preserve the original class distribution. This allowed the long-tailed characteristics of the real traffic to be retained while reducing computational cost.

The original dataset contains **85 features** and **39,603,674 records**. After preprocessing, it was reduced to **21 predictive features**. Instead of directly discarding temporal information, the timestamp components were reconstructed into initial and final flow-level references from the `ts_*` and `te_*` fields, then converted into the compact feature `td`, which summarizes flow duration. This preserved temporal information while reducing redundancy.

An exploratory analysis of the encoded feature space was used to examine cardinality and category distributions. In addition, the original description of LITNET-2020 indicates that part of the dataset was extended with custom attributes specifically designed for attack identification and attack type recognition. Therefore, variables such as `icmp_dst_ip_b`, `icmp_src_ip`, `udp_dst_p`, `tcp_f_s`, `tcp_f_n_a`, `tcp_f_n_f`, `tcp_f_n_r`, `tcp_f_n_p`, `tcp_f_n_u`, `tcp_dst_p`, `tcp_src_dst_f_s`, `tcp_src_tftp`, `tcp_src_kerb`, `tcp_src_rpc`, `tcp_dst_p_src`, and `smtp_dst` were removed because they behave as descriptive attack indicators rather than general traffic features. The binary target `attack_a` was also removed, together with constant, redundant, and leakage-prone variables. In total, **64 columns** were discarded. The final representation retained traffic descriptors related to duration, addressing, ports, protocol, flags, and packet- and byte-level statistics, yielding a more compact and less leakage-prone feature set for training.

B. Data Preprocessing

After feature selection, the retained variables were prepared for multiclass training through dataset partitioning, categorical encoding, and feature normalization. To avoid information leakage, all data-dependent transformations were fitted only on the training subset and then applied unchanged to the validation and test subsets.

1) Dataset Partitioning: The dataset was divided into stratified training, validation, and test subsets of **70%**, **15%**, and **15%**, respectively, preserving the original class distribution under severe multiclass imbalance.

A **stratified 5-fold cross-validation** procedure was also conducted to assess statistical robustness. In each fold, approximately **80%** of the data were used for training and **20%** for validation, with preprocessing fitted only on the corresponding training partition.

After preprocessing, the feature matrices were converted into PyTorch tensors with shape $(n_{\text{samples}}, 1, n_{\text{features}})$ for the convolutional model.

2) Categorical Encoding: Because LITNET-2020 contains high-cardinality categorical attributes, categorical variables were encoded using a CatBoost-based target encoding strategy. This approach preserves the statistical relationship between category values and target labels while avoiding the dimensional growth associated with one-hot encoding, particularly for source and destination addresses.

In general, the encoded value for a category c is defined as

$$\hat{y}_c = \frac{\sum_{i \in D_c} y_i}{|D_c|} \tag{1}$$

where $\hat{y}_c$ is the encoded value for category c, D_c is the subset of samples with category c, y_i is the target value of sample i, and $|D_c|$ is the number of such samples.

This strategy replaces each categorical value with a compact target-informed numerical representation. In this study, the category mappings were estimated using only the training data of each split and then applied unchanged to the corresponding validation and test subsets to reduce target leakage.

3) Feature Normalization: After categorical encoding, continuous features were normalized using min–max scaling:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{2}$$

where $\min(X)$ and $\max(X)$ denote the minimum and maximum observed values of feature X. This step reduces scale dominance and improves numerical stability during training. The scaling parameters were estimated only from the training subset and then applied to the validation and test subsets to avoid information leakage.

Algorithm 1 Preprocessing pipeline

Require: Raw dataset D
Ensure: Preprocessed train, validation, and test tensors

1: Load D
2: Remove irrelevant, redundant, constant, and leakage-prone columns
3: Encode multiclass labels as class indices
4: Identify categorical features C and continuous features N
5: Perform a stratified split into train, validation, and test subsets
6: Fit the CatBoost-based encoder on C using only the training subset
7: Apply the encoder to all subsets
8: Fit the min–max scaler on N using only the training subset
9: Apply the scaler to all subsets
10: Convert the subsets into PyTorch tensors of shape $(n_{\text{samples}}, 1, n_{\text{features}})$

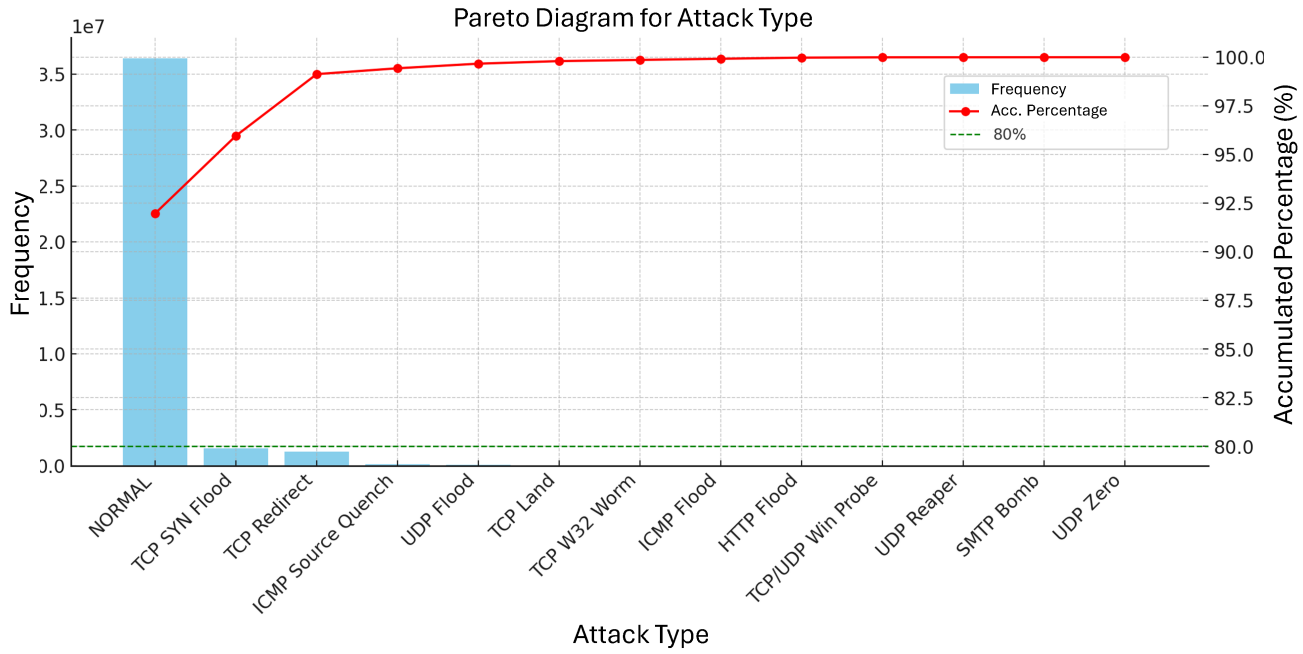

Fig. 1. Class distribution in LITNET-2020.

C. H.A.L.C.CO.N Architecture

H.A.L.C.CO.N is a convolutional multiclass intrusion detection architecture that incorporates hierarchical attention to improve feature discrimination in real-world flow-based traffic. It combines convolutional feature extraction, batch normalization, max-pooling, and attention-based feature reweighting in a unified design adapted to LITNET-2020.

Given the severe class imbalance, heterogeneous feature distributions, and high-cardinality categorical attributes of real network traffic, the architecture operates together with the preprocessing and optimization strategies described earlier, particularly CatBoost-based encoding and EQLv2-based learning. In this setting, the network learns progressively richer latent representations while hierarchical attention emphasizes the most informative patterns.

1) Network Structure: H.A.L.C.CO.N consists of three convolutional blocks followed by a fully connected classification layer. Each block includes a one-dimensional convolution, ReLU activation, max-pooling, batch normalization, and single-head hierarchical attention.

The first block uses 64 filters with kernel size 3 and max-pooling of size 4. The second uses 128 filters with kernel size 5 and max-pooling of size 2. The third uses 256 filters with kernel size 5 and max-pooling of size 2. Each block is followed by an attention layer. The final representation is flattened and mapped to the 13 output classes by a fully connected layer.

A schematic representation is shown in Fig. 2.

2) Mathematical Formulation: Let $x^{(l-1)}$ denote the input to block l. Each convolutional block applies

$$z^{(l)} = \mathrm{Conv1D}(x^{(l-1)}; W^{(l)}, b^{(l)}), \tag{3}$$

$$a^{(l)} = \mathrm{ReLU}(z^{(l)}), \qquad p^{(l)} = \mathrm{MaxPool}(a^{(l)}), \tag{4}$$

followed by batch normalization,

$$\hat{p}^{(l)} = \frac{p^{(l)} - \mu^{(l)}}{\sqrt{\sigma^{(l)2} + \epsilon}}, \qquad y^{(l)} = \gamma^{(l)}\hat{p}^{(l)} + \beta^{(l)}. \tag{5}$$

Hierarchical attention is then applied as

$$\alpha^{(l)} = \tanh(W^{(l)}_{\mathrm{attn}} y^{(l)} + b^{(l)}_{\mathrm{attn}}), \qquad \tilde{y}^{(l)} = y^{(l)} \odot \alpha^{(l)}, \tag{6}$$

where $\odot$ denotes element-wise multiplication. After the final attention stage, the representation is flattened and classified through

$$\hat{y} = \mathrm{Softmax}(W^{(f)} h + b^{(f)}), \tag{7}$$

where h is the flattened latent representation.

Algorithm 2 Forward pass of the H.A.L.C.CO.N model

Require: Input tensor $x \in \mathbb{R}^{(\mathrm{batch\_size},1,\mathrm{input\_features})}$
Ensure: Predicted class probabilities
1: Apply three Conv1D–ReLU–MaxPool1D–BatchNorm1D–attention blocks
2: Flatten the final representation
3: Apply the fully connected layer and softmax
4: **return** class probabilities

D. Equalization Loss v2 for Imbalanced Multiclass Learning

LITNET-2020 exhibits a severe long-tailed distribution in which the `Normal` class dominates and several attack categories are sparsely represented. Under this imbalance, conventional loss functions tend to favor majority classes, often yielding high overall accuracy but poor minority-class recall. To mitigate this effect, H.A.L.C.CO.N was trained

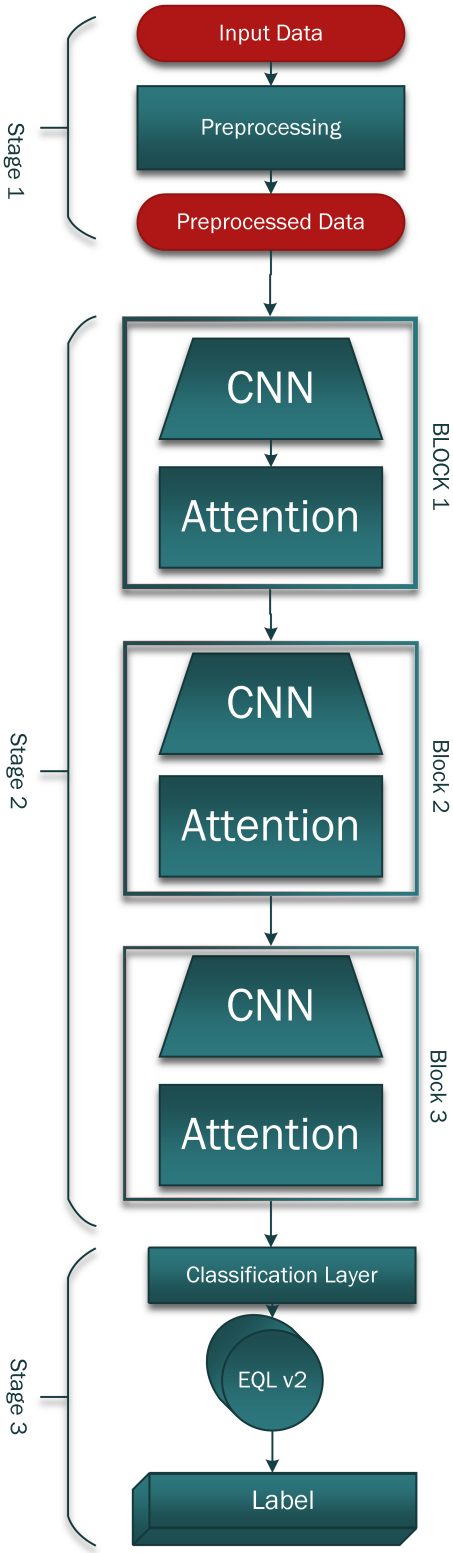

Fig. 2. Architecture of the proposed H.A.L.C.CO.N model.

with **Equalization Loss v2 (EQLv2)** adapted to multiclass intrusion detection.

EQLv2 dynamically reweights class-dependent gradient contributions to reduce majority-class dominance. Its main parameters were analyzed over approximate ranges of $\gamma \in [8.06, 15.97]$, $\mu \in [0.15, 0.44]$, and $\alpha \in [1.03, 2.99]$. The most favorable behavior was observed around ($\gamma = 12.0$, $\mu = 0.3$, $\alpha = 2.0$), which was retained in all reported experiments. In practice, EQLv2 improved rare-class discrimination while preserving stable learning for the dominant benign class.

E. Training Strategy

H.A.L.C.CO.N was trained with Adam using a learning rate of **0.0005**, a batch size of **256**, and up to **10 epochs**. The best checkpoint was selected according to validation loss, providing a practical balance between optimization stability and computational efficiency on LITNET-2020.

Training was monitored through validation loss, recall, and F1-score. Although accuracy was also reported, macro-level metrics were prioritized because the severe class imbalance of LITNET-2020 makes them more informative for minority-class evaluation.

F. Evaluation Metrics

Performance was evaluated using accuracy, macro-averaged recall, macro-averaged F1-score, and macro-averaged false positive rate (FPR). Given the severe class imbalance, macro-level metrics were prioritized over accuracy alone. Confusion matrices were also analyzed to examine class-wise behavior, identify dominant confusion patterns, and assess minority-class separability. Test loss was additionally monitored as an indicator of generalization.

1) Analysis Excluding the Majority Class: Because the `Normal` class dominates LITNET-2020, an additional analysis was performed to better examine minority attack categories. This reduces the masking effect of the majority class in the confusion matrix and related metrics.

Let c_{maj} denote the majority class in the test set:

$$c_{\text{maj}} = \arg\max_{c} f_c, \tag{8}$$

where f_c is the number of test samples in class c. The filtered labels and predictions were then defined as

$$y_{\text{true}}^{(-\text{maj})} = \{y_i \mid y_i \neq c_{\text{maj}}\}, \qquad y_{\text{pred}}^{(-\text{maj})} = \{\hat{y}_i \mid y_i \neq c_{\text{maj}}\}. \tag{9}$$

Using these sets, a reduced confusion matrix was computed:

$$\text{CM}^{(-\text{maj})} = \text{confusion\_matrix}\big(y_{\text{true}}^{(-\text{maj})}, y_{\text{pred}}^{(-\text{maj})}\big). \tag{10}$$

Class-wise false positive rates were also calculated as

$$\text{FPR}_c = \frac{FP_c}{FP_c + TN_c}, \tag{11}$$

with weighted aggregation given by

$$\text{FPR}_{\text{weighted}} = \sum_{c=1}^{C} \text{FPR}_c \cdot \frac{f_c}{\sum_{j=1}^{C} f_j}. \tag{12}$$

This complementary analysis provides a clearer view of performance on underrepresented attack categories.

TABLE I
COMPARISON WITH CLASSICAL BASELINES ON LITNET-2020

Model	Acc.	Rec.	FPR	F1
Proposed Model	0.999559	0.999559	0.00013	0.999523
Logistic Regression	0.960000	0.960000	0.04000	0.940000
KNN	0.970000	0.970000	0.03000	0.970000
SVM	0.960000	0.960000	0.04000	0.940000

G. Implementation Details

The model was implemented in **PyTorch** using **Python 3.12** and **PyTorch 2.3.0+cu118**. Experiments were conducted on a system with an **NVIDIA GPU with 6 GB of memory**, and the complete training and evaluation process required approximately **8 hours**.

IV. RESULTS AND COMPARATIVE EVALUATION

A. Performance on Real-World Data

The proposed **H.A.L.C.CO.N** model achieved strong performance on the real-world **LITNET-2020** dataset. On the test set, it reached an **accuracy** of **99.996%**, a **recall** of **99.997%**, and a **false positive rate (FPR)** of **0.0135%**, indicating robust multiclass discrimination under realistic and severely imbalanced traffic conditions.

Class-wise analysis showed near-perfect discrimination for dominant categories such as `Normal` and `tcp_syn_f`, while performance remained more challenging for severely underrepresented classes, including `smtp_b` and `udp_0`. The class-wise FPR remained close to zero for dominant categories and increased mainly in minority classes, consistent with the long-tailed distribution of LITNET-2020.

B. Comparison with Classical Machine Learning Baselines

To provide a direct reference on the same multiclass task, multinomial Logistic Regression, KNN, and SVM were evaluated on LITNET-2020 under the same real-world setting.

As shown in Table I, H.A.L.C.CO.N achieved the best overall performance, reaching an accuracy of 0.999559, a recall of 0.999559, an F1-score of 0.999523, and an FPR of 0.00013. KNN was the strongest classical baseline, whereas Logistic Regression and SVM showed the weakest results, particularly because of their higher false positive rates.

These results confirm that, under severe multiclass imbalance and real-world flow-based traffic conditions, the proposed architecture is more effective than classical machine learning baselines, especially in terms of false alarm control.

C. Cross-Validation Analysis

To further assess robustness under the severe class imbalance of LITNET-2020, a stratified 5-fold cross-validation analysis was performed. Table II reports the mean and standard deviation of the main evaluation metrics across folds.

TABLE II
STRATIFIED 5-FOLD CROSS-VALIDATION RESULTS FOR H.A.L.C.CO.N ON THE 10% LITNET

Metric	Mean $\pm$ Std
Accuracy	0.99965 ± 0.00004
Precision$_{macro}$	0.9510 ± 0.0302
Recall$_{macro}$	0.9134 ± 0.0280
F1-score$_{macro}$	0.9225 ± 0.0173
FPR$_{macro}$	0.000199 ± 0.000064
Precision$_{weighted}$	0.99965 ± 0.00004
Recall$_{weighted}$	0.99965 ± 0.00004
F1-score$_{weighted}$	0.99964 ± 0.00004

D. Comparison with State-of-the-Art IDS Models

Table III compares H.A.L.C.CO.N with representative IDS models previously evaluated on benchmark datasets such as NSL-KDD and UNSW-NB15, including both classical machine learning and deep learning approaches [5], [10]–[14], [20].

Previous IDS models have reported competitive results on controlled benchmark datasets [5], [10]–[14], [20]. However, many of these datasets do not fully capture the complexity, heterogeneity, and class imbalance of real-world traffic. Under this broader perspective, H.A.L.C.CO.N remained highly competitive while being evaluated on real flow-based traffic. Although cross-study comparisons must be interpreted cautiously because of differences in datasets, preprocessing, class definitions, and evaluation protocols, the results suggest that the combination of hierarchical attention, CatBoost-based encoding, and EQLv2 is effective for real-world multiclass intrusion detection.

E. Confusion Matrix Analysis

Fig. 3 shows the confusion matrix on the LITNET-2020 test set. The matrix exhibits a strong diagonal structure, indicating consistent class-wise discrimination across most traffic categories. The dominant `Normal` class was correctly classified in **546,162** of **546,285** instances, corresponding to a recall of approximately **99.98%**. Several attack classes, including `icmp_f`, `tcp_udp_win_p`, and `udp_reaper_w`, achieved perfect recall, while `icmp_smf`, `tcp_land`, `tcp_red_w`, and `tcp_syn_f` remained above **99.8%**.

Residual errors were concentrated in severely underrepresented classes, particularly `smtp_b` and `udp_0`, with recalls of approximately **36.36%** and **57.14%**, respectively. Most errors corresponded to misclassification into the `Normal` class rather than confusion among attack categories, indicating that the main challenge remains the separation of rare attacks from dominant benign traffic. Misclassifications from `Normal` into attack classes were sparse, consistent with the low overall false positive rate. These results confirm robust performance across most traffic categories while highlighting the persistent difficulty of extremely rare attacks.

TABLE III
PERFORMANCE COMPARISON OF THE PROPOSED MODEL ON LITNET-2020 AGAINST PRIOR IDS MODELS EVALUATED ON BENCHMARK DATASETS

Model	Dataset	Accuracy (%)	Recall (%)	FPR (%)
SVM	NSL-KDD	69.52	–	–
SVM	UNSW-NB15	74.80	83.71	7.73
RF	UNSW-NB15	84.59	92.24	3.01
AdaBoost	UNSW-NB15	73.19	91.13	22.11
HAST-IDS	NSL-KDD	93.27	95.85	–
HAST-IDS	UNSW-NB15	80.03	93.65	9.60
CNN-BiLSTM	NSL-KDD	99.22	98.88	0.43
CNN-BiLSTM	UNSW-NB15	82.08	92.50	6.09
LuNet	NSL-KDD	99.14	99.02	0.61
LuNet	UNSW-NB15	85.35	97.73	2.89
Pelican	NSL-KDD	99.21	99.13	0.65
Pelican	UNSW-NB15	86.64	97.75	1.30
CANET	NSL-KDD	99.77	99.72	0.18
CANET	UNSW-NB15	89.39	98.93	0.87
Proposed Model	**LITNET-2020**	**99.99**	**99.99**	**0.0135**

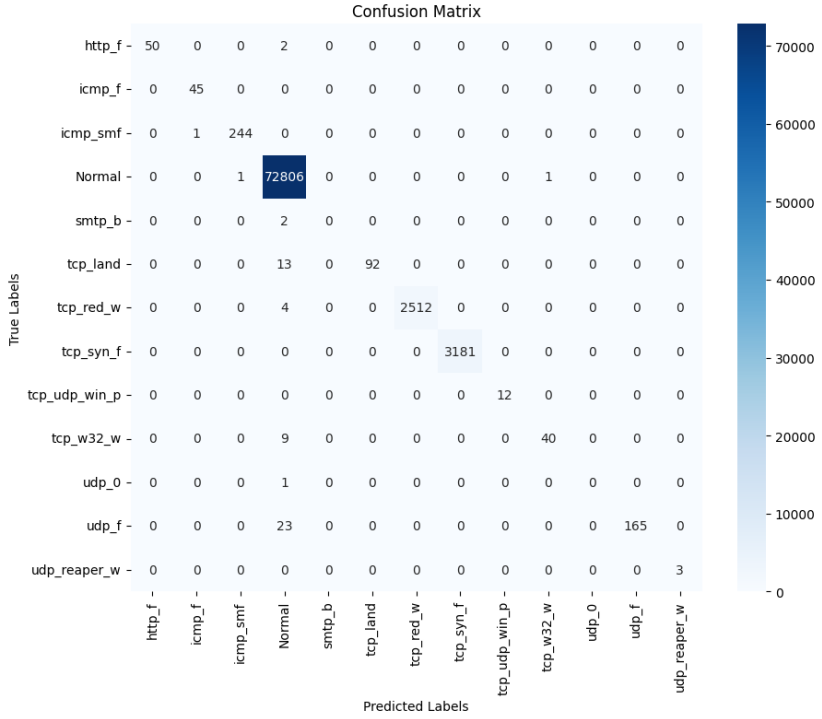

Fig. 3. Confusion matrix of H.A.L.C.CO.N on the LITNET-2020 test set using logarithmic color scaling.

F. Ablation Study

A partial ablation study was conducted using a 2×2 design in which the categorical encoding strategy and the hierarchical attention module were varied, while EQLv2 was kept fixed. Two encoding strategies were evaluated, CatBoost-based encoding and standard label encoding, each with and without attention.

Table IV summarizes the results. The full H.A.L.C.CO.N configuration achieved the best macro-F1 score (0.9076), with an accuracy of 0.9997 and a macro-FPR of 1.35×10^{-4}. The encoding strategy had the largest effect: replacing CatBoost with label encoding reduced macro-F1 from 0.9076 to 0.6224 with attention and from 0.9065 to 0.6612 without attention, indicating that CatBoost substantially improves class separability.

By contrast, hierarchical attention provided a smaller effect. Under CatBoost-based encoding, attention slightly improved macro-F1 from 0.9065 to 0.9076, whereas under label encoding it reduced performance. In terms of efficiency, the CatBoost-without-attention variant achieved the best macro-FPR (1.21×10^{-4}) and the lowest latency and memory usage, making it an attractive alternative when computational cost

TABLE IV
PARTIAL ABLATION RESULTS FOR H.A.L.C.CO.N WITH EQLV2 FIXED

Encoding	Attn.	Acc.	F1-macro	FPR-macro	Latency	Memory
CatBoost	Yes	0.9997	**0.9076**	0.000135	0.00320	1451.0
Label	Yes	0.9968	0.6224	0.001797	0.00261	1449.1
CatBoost	No	0.9997	0.9065	0.000121	0.00212	1435.4
Label	No	0.9968	0.6612	0.001357	0.00226	1435.9

is prioritized. Overall, the results indicate that CatBoost-based encoding is the main source of performance gain, while hierarchical attention provides a smaller complementary refinement.

G. Effect of EQLv2 Hyperparameters on Model Performance

A sensitivity analysis was conducted to examine the effect of the main EQLv2 hyperparameters on H.A.L.C.CO.N under the severe class imbalance of LITNET-2020. The explored ranges were approximately $\gamma \in [8.06, 15.97]$, $\mu \in [0.15, 0.44]$, and $\alpha \in [1.03, 2.99]$.

Functionally, α controls the magnitude of the positive reweighting term, γ regulates the steepness of the logistic mapping used in the dynamic weighting mechanism, and μ determines the midpoint at which this reweighting becomes more influential. Together, these parameters govern how strongly EQLv2 counteracts majority-class dominance and supports minority-class learning.

The results showed clear sensitivity to parameter selection. Weak reweighting reduced minority-class sensitivity, whereas overly aggressive reweighting increased performance variability. The most favorable behavior was observed around $(\gamma = 12.0,\ \mu = 0.3,\ \alpha = 2.0)$, suggesting that EQLv2 is most effective when minority-class support is improved without destabilizing global optimization.

Overall, these findings support the usefulness of EQLv2 for real-world multiclass intrusion detection and show that its contribution depends on principled parameter calibration.

H. Cross-Dataset Evaluation on UNSW-NB15

To assess the robustness of H.A.L.C.CO.N beyond LITNET-2020, an additional evaluation was conducted on UNSW-NB15. For this experiment, the architecture was adapted only to the input dimensionality and number of classes required by the dataset, while preserving its core components: convolutional hierarchy, hierarchical attention, CatBoost-based encoding, and EQLv2 loss. The categorical features `proto`, `service`, and `state` were encoded with CatBoost-based encoding, labels were transformed with label encoding, and min–max normalization was applied using training-set statistics only.

On UNSW-NB15, H.A.L.C.CO.N achieved a test accuracy of **72.16%**, macro-precision of **47.14%**, macro-recall of **47.77%**, macro-F1 of **44.98%**, and macro false positive rate of **3.06%**. These results are lower than those obtained on LITNET-2020, indicating degraded performance under a different dataset configuration and traffic distribution.

A plausible explanation is that the UNSW-NB15 representation used here provides a more restricted attribute set than LITNET-2020 and than the complete raw UNSW-NB15 traffic files. Thus, the lower performance may reflect not only dataset shift and attack heterogeneity, but also limitations of the available feature representation. Although this experiment does not constitute a full domain-transfer study, it provides evidence that H.A.L.C.CO.N was also evaluated on an independent IDS benchmark.

I. Discussion

The results show that H.A.L.C.CO.N is effective for multiclass intrusion detection under real-world traffic conditions. On LITNET-2020, the model achieved strong performance together with a very low false positive rate, which is especially relevant in operational cybersecurity settings. The confusion matrix showed that most remaining errors were concentrated in severely underrepresented attack classes, often confused with `Normal`, indicating that the main challenge lies in preserving discriminative boundaries for rare attacks under extreme imbalance.

The ablation study indicates that CatBoost-based encoding contributes the largest gain, highlighting the importance of robust categorical representation for real-world IDS data. Hierarchical attention provides a complementary improvement, while EQLv2 increases minority-class sensitivity when properly calibrated, with the best behavior observed around $(\gamma = 12.0,\ \mu = 0.3,\ \alpha = 2.0)$. The additional results on UNSW-NB15 further suggest that performance depends not only on the architecture, but also on the richness and contextual completeness of the available feature set.

J. Limitations and Threats to Validity

Although H.A.L.C.CO.N achieved strong performance on LITNET-2020, several limitations should be acknowledged.

1) Internal Validity: The reported results depend on the preprocessing and experimental pipeline adopted in this study, including feature cleaning, categorical encoding, normalization, and stratified partitioning. To reduce information leakage, leakage-prone fields were removed and all data-dependent preprocessing steps were fitted exclusively on the training partition of each split and then applied unchanged to validation and test data, including within cross-validation folds. Nevertheless, as in any supervised pipeline on structured traffic data, some dataset-specific correlations may still favor classification performance.

A second threat arises from the severe multiclass imbalance of LITNET-2020. Under this condition, global metrics such as accuracy may be dominated by the majority class and may not fully reflect minority-class performance. For this reason, the analysis also included macro-averaged precision, recall, F1-score, false positive rate, per-class results, confusion matrices, and stratified 5-fold cross-validation. Even so, minority classes with very limited support remain difficult to estimate with high statistical confidence.

2) External Validity: External validity is bounded by the characteristics of the LITNET-2020 environment. Although real-world traffic provides stronger ecological validity than synthetic or outdated benchmark datasets, the results cannot be

directly generalized to all operational networks. Differences in traffic policies, infrastructure, temporal behavior, user activity, attack prevalence, and protocol distributions may affect model performance.

In addition, the proposed model was evaluated on a fixed feature space derived from a specific environment. Therefore, the reported results should be interpreted as evidence of effectiveness on LITNET-2020 rather than proof of universal generalization. Further validation on independent datasets and cross-dataset scenarios is needed to assess robustness under domain shift.

3) Scope of the Present Study: This study was designed to evaluate the contribution of hierarchical attention, convolutional representation learning, CatBoost-based encoding, and EQLv2 in a controlled multiclass intrusion detection setting. It does not address online adaptation, continual learning, concept drift, adversarial robustness, or deployment-time constraints in production-scale IDS environments, which remain relevant directions for future work.

V. Conclusion and Future Work

A. Conclusions

This study evaluated **H.A.L.C.CO.N** for multiclass intrusion detection on the real-world LITNET-2020 dataset. The results show that the proposed model achieves strong performance under severe class imbalance and heterogeneous traffic conditions, combining high accuracy and recall with a very low false positive rate. The analyses further indicate that CatBoost-based encoding is the main source of performance gain, hierarchical attention provides a smaller complementary improvement, and EQLv2 contributes to minority-class sensitivity when appropriately calibrated. Although severely underrepresented attack categories remain challenging, the overall results support the effectiveness of combining convolutional learning, hierarchical attention, CatBoost-based encoding, and EQLv2 for real-world multiclass intrusion detection.

B. Future Work

Future work should focus on improving the detection of extremely rare attack classes, validating the model across additional traffic environments, and further reducing inference cost for real-time deployment. Extensions of the attention mechanism and adaptive learning strategies may also improve robustness under evolving traffic and threat conditions.

Acknowledgements

The first author sincerely acknowledges the directors and professors of the **División Académica de Ciencias y Tecnologías de la Información (DACYTI)** for their guidance and encouragement throughout this research. The first author also gratefully acknowledges **CONACYT** for the scholarship support that made this research possible.

The first author also expresses his gratitude to **Weiying Qu**, **Ma YongHai**, and **Luo ShaoQing** for their institutional and professional support, and for facilitating the time and working conditions that enabled the completion of his graduate studies and this research work.

References

[1] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, "A new intrusion detection system based on knn classification algorithm in wireless sensor network," *Journal of Electrical and Computer Engineering*, vol. 2014, p. 240217, 2014. [Online]. Available: https://doi.org/10.1155/2014/240217

[2] S. K. Wagh, V. K. Pachghare, and S. R. Kolhe, "Survey on intrusion detection system using machine learning techniques," *International Journal of Computer Applications*, vol. 78, 2013. [Online]. Available: https://doi.org/10.5120/13608-1412

[3] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41 525–41 550, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2895334

[4] J. Yan, D. Jin, C. W. Lee, and P. Liu, "A comparative study of off-line deep learning based network intrusion detection," in *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2018, pp. 299–304. [Online]. Available: https://doi.org/10.1109/ICUFN.2018.8436774

[5] K. Ren, S. Yuan, C. Zhang, Y. Shi, and Z. Huang, "CANET: A hierarchical CNN-attention model for network intrusion detection," *Computer Communications*, vol. 205, pp. 170–181, 2023. [Online]. Available: https://doi.org/10.1016/j.comcom.2023.04.018

[6] J. Tan, X. Lu, G. Zhang, C. Yin, and Q. Li, "Equalization loss v2: A new gradient balance approach for long-tailed object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 1685–1694. [Online]. Available: https://doi.org/10.1109/CVPR46437.2021.00173

[7] N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/MilCIS.2015.7348942

[8] R. Damasevicius, A. Venckauskas, S. Grigaliunas, J. Toldinas, N. Morkevicius, T. Aleliunas, and P. Smuikys, "LITNET-2020: An annotated real-world network flow dataset for network intrusion detection," *Electronics*, vol. 9, no. 5, p. 800, 2020. [Online]. Available: https://doi.org/10.3390/electronics9050800

[9] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, 2019. [Online]. Available: https://doi.org/10.1186/s42400-019-0038-7

[10] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33 789–33 795, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2018.2841987

[11] W. Hu, J. Gao, Y. Wang, O. Wu, and S. Maybank, "Online adaboost-based parameterized methods for dynamic distributed network intrusion detection," *IEEE Transactions on Cybernetics*, vol. 44, no. 1, pp. 66–82, 2014. [Online]. Available: https://doi.org/10.1109/TCYB.2013.2247592

[12] W. Wang, Y. Sheng, J. Wang, X. Zeng, X. Ye, Y. Huang, and M. Zhu, "HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection," *IEEE Access*, vol. 6, pp. 1792–1806, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2780250

[13] J. Sinha and M. Manollas, "Efficient deep CNN-BiLSTM model for network intrusion detection," in *Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition*, 2020, pp. 223–231. [Online]. Available: https://doi.org/10.1145/3430199.3430224

[14] P. Wu, H. Guo, and N. Moustafa, "Pelican: A deep residual network for network intrusion detection," in *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 2020, pp. 55–62. [Online]. Available: https://doi.org/10.1109/DSN-W50199.2020.00018

[15] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "A comprehensive review of attention mechanisms in deep learning," *arXiv preprint arXiv:1904.02874*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.1904.02874

[16] J. Kim, J. Kim, and H. K. Kim, "Attention-based recurrent neural network for network intrusion detection," in *2018 International Conference on Big Data and Smart Computing (BigComp)*, 2018, pp. 313–316. [Online]. Available: https://doi.org/10.1109/BIGCOMP.2018.00055

[17] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *2010 IEEE Symposium on Security and Privacy*, 2010, pp. 305–316. [Online]. Available: https://doi.org/10.1109/SP.2010.25
[18] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *2019 International Carnahan Conference on Security Technology (ICCST)*, 2019, pp. 1–8. [Online]. Available: https://doi.org/10.1109/CCST.2019.8888419
[19] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in Neural Information Processing Systems*, vol. 31, 2018. [Online]. Available: https://doi.org/10.48550/arXiv.1810.11363
[20] P. Wu and H. Guo, "LuNet: a deep neural network for network intrusion detection," in *2019 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2019, pp. 617–624. [Online]. Available: https://doi.org/10.1109/SSCI44817.2019.9003126

Rodolfo Martinez Cadena received the bachelor's degree in Electronics and Communications Engineering from Universidad Olmeca in 2003 and the M.Sc. degree in Computational Sciences from Universidad Juárez Autónoma de Tabasco. He has led AI-driven initiatives at *SEIENER* and has experience in applied technology projects involving data analysis, intelligent systems, and operational decision support. His research interests include artificial intelligence, geophysical exploration, real-time data analysis, brain-computer interfaces, and the Internet of Things (IoT).

José Adán Hernández Nolasco received the bachelor's degree in electronic and communications engineering from the Autonomous University of Nuevo León, in 1996, the M.Sc. degree in electronic engineering (telecommunications) from the Monterrey Institute of Technology and Higher Education, in 2003, and the Ph.D. degree in optics from the National Institute for Astrophysics, Optics and Electronics, in 2012. He has been a Research Professor with the Universidad Juárez Autónoma de Tabasco, for 25 years. He has authored or coauthored over 25 publications in the areas of ambient intelligence and AI applications, and over 30 participations in conferences. His research interests include artificial intelligence, fuzzy logic, IoT, and optics.

Noel Zacarias-Morales is a graduate of the Ph.D. in Computer Science at the Academic Division of Information Sciences and Technologies at the Universidad Juárez Autónoma de Tabasco. He obtained his Master's degree in Information Technology Management at the Universidad Juárez Autónoma de Tabasco in 2019. His research interests are the development of models based on deep learning (artificial neural networks) applied to signal analysis and processing for prediction and classification.