

A TWO-TIER HYBRID INTRUSION DETECTION SYSTEM FOR IoT/MQTT NETWORKS BASED ON MACHINE LEARNING ENSEMBLE AND FUZZY INFERENCE SYSTEM

Nguyen Thi Hong Hoa*, Pham Dinh Phong

University of Transport and Communications

ARTICLE INFO		ABSTRACT
Received:	25/5/2026	This paper proposes a two-stage hybrid intrusion detection system for IoT/MQTT networks. The first stage employs an Ensemble Stacking model combining Multi-Layer Perceptron (MLP), Random Forest, and eXtreme Gradient Boosting (XGBoost), with Logistic Regression as the meta-learner. The second stage applies a Fuzzy Inference System (FIS) to selectively re-examine low-confidence legitimate predictions. On the MQTTEEB-D dataset (60/20/20 split, SMOTE on training set only), the baseline Ensemble achieves the values of F1-Weighted, F1-Macro, Legit-F1, and Macro AP of 0.9450, 0.9152, 0.6866, and 0.9497, respectively. The proposed Hybrid Ensemble-FIS further improves these values to 0.9465, 0.9174, 0.6888, 0.9478, recovering 193 brute-force samples previously misclassified as legitimate and raising the recall of this class from 89% to 92%. The paper also presents a Precision–Recall analysis instead of ROC curves, which better fits the imbalanced data context. The proposed method enhances intrusion detection in IoT/MQTT networks while preserving interpretability through its Mamdani rule-based design.
Revised:	29/6/2026	
Published:	30/6/2026	
KEYWORDS		
MQTT		
IoT		
Intrusion detection		
Fuzzy inference system		
Machine learning ensemble		

HỆ THỐNG PHÁT HIỆN XÂM NHẬP LẠI HAI TẦNG CHO MẠNG IoT/MQTT DỰA TRÊN ENSEMBLE HỌC MÁY VÀ HỆ SUY DIỄN MỜ

Nguyễn Thị Hồng Hoa*, Phạm Đình Phong

Trường Đại học Giao thông vận tải

THÔNG TIN BÀI BÁO		TÓM TẮT
Ngày nhận bài:	25/5/2026	Bài báo đề xuất một hệ thống phát hiện xâm nhập lại hai tầng cho mạng IoT sử dụng giao thức MQTT. Tầng 1 áp dụng mô hình Ensemble Stacking kết hợp Multi-Layer Perceptron (MLP), Random Forest và eXtreme Gradient Boosting (XGBoost), trong đó Logistic Regression đóng vai trò meta-learner. Tầng 2 sử dụng Hệ suy diễn mờ (Fuzzy Inference System – FIS) để xem xét lại có chọn lọc các mẫu được dự đoán là hợp lệ nhưng độ tin cậy chưa cao. Thử nghiệm trên bộ dữ liệu MQTTEEB-D (chia 60/20/20, SMOTE chỉ áp dụng cho tập huấn luyện) cho thấy mô hình Ensemble cơ sở đạt F1-Weighted, F1-Macro, Legit-F1 và Macro AP lần lượt là 0,9450, 0,9152, 0,6866 và 0,9497. Mô hình lai Ensemble-FIS cải thiện các chỉ số này lần lượt là 0,9465, 0,9174, 0,6888 và 0,9478, đồng thời khôi phục 193 mẫu bruteforce bị phân loại nhầm, nâng Recall lớp này từ 89% lên 92%. Bài báo sử dụng phân tích Precision–Recall thay cho ROC để phù hợp với bối cảnh dữ liệu mất cân bằng. Phương pháp đề xuất nâng cao hiệu quả phát hiện xâm nhập trong mạng IoT/MQTT, đồng thời đảm bảo tính diễn giải nhờ hệ luật Mamdani.
Ngày hoàn thiện:	29/6/2026	
Ngày đăng:	30/6/2026	
TỪ KHÓA		
MQTT		
IoT		
Phát hiện xâm nhập		
Hệ suy diễn mờ		
Ensemble học máy		

DOI: <https://doi.org/10.34238/tnu-jst.15899>

* Corresponding author. Email: hoanth@utc.edu.vn

1. Giới thiệu

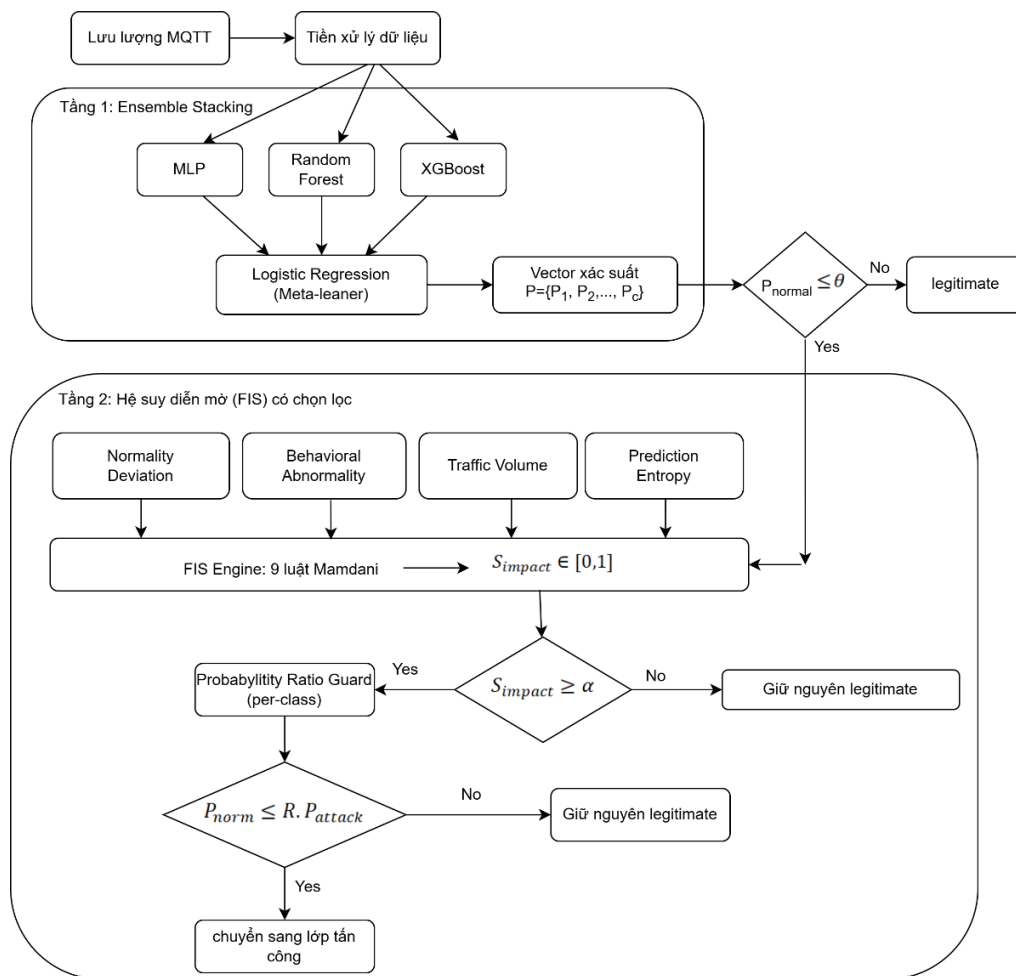
Sự phát triển nhanh chóng của các thiết bị Internet vạn vật (Internet of Things - IoT) trong những năm gần đây, do đặc điểm tài nguyên hạn chế và băng thông thấp nên cần sử dụng các giao thức truyền tin nhẹ. MQTT (Message Queuing Telemetry Transport) là một chuẩn mở của OASIS và đã được công nhận là tiêu chuẩn quốc tế ISO/IEC 20922, được thiết kế theo mô hình publish/subscribe phù hợp với các thiết bị IoT [1], [2]. Đặc tính nhẹ này làm cho MQTT thiếu các cơ chế bảo mật ở lớp ứng dụng [3], dễ bị tấn công vét cạn (bruteforce), từ chối dịch vụ (DoS), tấn công làm tràn publish (flood), gói tin dị dạng (malformed) và tấn công Slow-DoS (slowite) [4], [5]. Từ đó đặt ra thách thức trong việc nghiên cứu một hệ thống phát hiện xâm nhập (IDS) chuyên biệt cho MQTT cho kiến trúc bảo mật IoT hiện đại. Trong đó có vấn đề mất cân bằng lớp là một hạn chế của IDS dựa trên học máy [6], [7]. Trong dữ liệu MQTT, lưu lượng hợp lệ (*legitimate*) thường chiếm tỉ lệ thấp hơn nhiều so với tổng lưu lượng tấn công, khiến các mô hình nhận diện sẽ nhận diện đúng thiên về lớp đa số và có recall rất thấp với *legitimate*. Ngoài ra, hai lớp *legitimate* và *bruteforce* có nhiều đặc trưng tương đồng do *bruteforce* mô phỏng các phiên đăng nhập hợp lệ liên tiếp, làm cho việc phân biệt chúng trở nên khó hơn. Bên cạnh đó, đường cong ROC và AUC thường được dùng để đánh giá IDS có thể gây cảm nhận sai lệch lạc quan trong bối cảnh mất cân bằng do thành phần âm tính thật (True Negative-TN) lớn chiếm đa số. Bài báo số [8] đã chỉ ra rằng đường cong Precision-Recall (PR) và Average Precision (AP) cung cấp đánh giá thông tin hơn cho lớp thiểu số. Bên cạnh đó bài báo [9] nghiên cứu các biến thể mở rộng như precision/recall reject curves cho phép phân tích sâu hơn về độ tin cậy phân loại theo ngưỡng từ chối – đây là một hướng phát triển bổ sung cho thang đo PR khi cần đánh giá rủi ro quyết định trong các hệ IDS có cơ chế lựa chọn từ chối.

Các hệ thống phát hiện xâm nhập hiện đại đã được nghiên cứu dựa trên một mô hình học máy như MLP, CNN hoặc LSTM [10], [11]. Mặc dù các mô hình này đạt hiệu suất tốt trên lớp đa số nhưng lại kém hiệu quả với các lớp thiểu số, đặc biệt là phân biệt giữa lưu lượng hợp lệ và một số dạng tấn công có đặc trưng tương đồng như *bruteforce*. Thêm nữa, trong môi trường an ninh mạng thực tế, một hệ IDS cần có độ tin cậy và tính giải thích với các nguy cơ mất an toàn thông tin mà các mô hình học sâu (deep learning) chưa đáp ứng được. Nhiều nghiên cứu trước đây đã sử dụng bộ dữ liệu MQTT-IoT-IDS2020 để đánh giá các thuật toán học máy truyền thống như Cây quyết định, Rừng ngẫu nhiên và Bayes đơn giản trong các bài toán phát hiện xâm nhập nhị phân và đa lớp [12]. Một số nghiên cứu khác cũng khai thác MLP nhằm phát hiện các dạng tấn công MQTT như malformed CONNECT hay slow ACK [13]. Tuy nhiên, trong điều kiện dữ liệu mất cân bằng mạnh, hiệu quả nhận diện lớp lưu lượng hợp lệ vẫn còn hạn chế. Để cải thiện vấn đề này, nhiều phương pháp dựa trên Rừng ngẫu nhiên kết hợp cơ chế gán trọng số lớp [14] đã được áp dụng trên các bộ dữ liệu IDS như NSL-KDD và cho thấy khả năng cải thiện đáng kể đối với các lớp thiểu số. Gần đây, kỹ thuật Stacking với bộ siêu học viên (meta-learner) [15] cũng được sử dụng để kết hợp đầu ra từ nhiều mô hình cơ sở trong hệ thống IDS. Dù vậy, phần lớn các nghiên cứu hiện nay vẫn tập trung vào kiến trúc kết hợp mà chưa tích hợp cơ chế suy luận mờ ở tầng hậu xử lý. Vì vậy, việc kết hợp kỹ thuật học kết hợp (Ensemble Learning) với hệ thống suy diễn mờ (Fuzzy Inference System – FIS) đang được xem là một hướng tiếp cận tiềm năng nhằm nâng cao khả năng xử lý các trường hợp phân loại không chắc chắn trong IDS [16], [17]. Bài báo [18] đã đề xuất khung FLAME kết hợp logic mờ với thuật toán tìm kiếm tối ưu Artificial Lizard nhằm nâng cao bảo mật IoT. Bài báo [19] lại sử dụng độ đo thông tin tương hỗ mờ Gaussian (Fuzzy-MIFS) để chọn lọc đặc trưng cho IDS-IoT, cải thiện độ chính xác từ 0,96 lên 0,986 trên kiến trúc CNN. Tổng hợp lại, có thể nhận thấy phần lớn các nghiên cứu hiện tại tập trung vào mô hình kết hợp đơn thuần mà chưa tích hợp cơ chế suy luận mờ ở tầng sau đó; các FIS hiện có thường yêu cầu ngưỡng cố định hoặc phụ thuộc nhãn ground-truth làm giảm tính khả thi khi triển khai thực tuyến và chưa có nhiều công trình đánh giá đầy đủ trên dữ liệu MQTT thu thập từ môi trường thực với phương pháp đánh giá phù hợp cho dữ liệu mất cân bằng với tính diễn giải cao.

Để giải quyết những hạn chế nêu trên, bài báo đề xuất một hệ thống phát hiện xâm nhập lai hai tầng kết hợp Ensemble Stacking và Hệ suy diễn mờ Mamdani cho mạng IoT/MQTT. Tầng 1 sử dụng Ensemble Stacking kết hợp ba mô hình hỗ trợ MLP, Random Forest [20] và XGBoost [21] với meta-learner Logistic Regression theo nguyên lý Stacked Generalization [15]. Tầng 2 sử dụng FIS Mamdani [22] dựa trên lý thuyết tập mờ của Zadeh [23] với bốn biến đầu vào (Normality Deviation, Behavioral Abnormality, Traffic Volume, Prediction Entropy) và chín luật mờ để xem xét lại có chọn lọc các mẫu được dự đoán là hợp lệ nhưng có độ tin cậy chưa đủ cao. Tại Tầng 2 việc thiết lập ngưỡng riêng cho từng lớp tấn công được xác định tự động từ phân phối dữ liệu thực tế tránh cơ chế của FIS truyền thống thường đề ngưỡng cố định. Tiếp theo sử dụng phân tích Precision-Recall và Average Precision per-class thay vì ROC/AUC để đảm bảo phản ánh trung thực hiệu suất trên lớp thiểu số. Phần tiếp theo của bài báo gồm: Mục 2 trình bày phương pháp nghiên cứu; mục 3 kết quả và thảo luận; mục 4 là kết luận.

2. Phương pháp nghiên cứu

Hệ thống đề xuất gồm hai tầng nối tiếp như mô tả ở Hình 1. Đầu vào là vector đặc trưng từ lưu lượng MQTT. Tầng 1 là Ensemble Stacking cho vector xác suất phân lớp. Nếu Tầng 1 dự đoán *legitimate* với độ tin cậy đủ cao, hệ thống trả kết quả ngay; ngược lại Tầng 2 FIS được kích hoạt để tích hợp các đặc trưng mờ và quyết định có phân loại sang một lớp attack hay không.



Hình 1. Sơ đồ kiến trúc hệ thống Hybrid Ensemble-FIS hai tầng

2.1. Tiền xử lý và đặc trưng

Các đặc trưng của MQTT (tcp_flags , $mqtt_conack_flags$, $mqtt_conflags$, $mqtt_hdrflags$) dưới dạng số thập phân được chuyển sang số nguyên. Ba đặc trưng dẫn xuất được tạo thêm: $tcp_header_ratio = tcp_len / (Header_Length + 1)$, $time_delta_{sq} = (tcp_time_delta)^2$, và $kalive_{qos} = mqtt_kalive \times (mqtt_qos + 1)$. Không gian đặc trưng cuối cùng gồm 9 trường lấy ở mức gói, không sử dụng nội dung dữ liệu. Bộ dữ liệu MQTTEEB-D thực nghiệm được chia 60/20/20 cho train/test/holdout bằng phương pháp lấy mẫu phân tầng (random_state=42); việc chuẩn hóa dữ liệu được tính toán từ tập huấn luyện thô; cân bằng dữ liệu sử dụng SMOTE [24] áp dụng sau chuẩn hóa và chỉ áp dụng trên tập huấn luyện để tránh rò rỉ. Giá trị $k_neighbors=3$ đây là giá trị mặc định của thư viện imbalanced-learn [24] cho dữ liệu có lớp thiểu số quy mô vừa, trong tập huấn luyện, lớp *legitimate* có 17.557 mẫu nên với trường hợp $k_neighbors=1$ dễ gây overfit cục bộ, còn nếu $k_neighbors=5$ sẽ sinh ra mẫu xa lớp gốc. Do đó Giá trị $k_neighbors = 3$ cân bằng giữa độ đa dạng của mẫu tổng hợp và tính đại diện cục bộ phù hợp với kích thước lớp thiểu số quy mô vừa.

2.2. Ensemble Stacking

Mô hình ở Tầng 1 được triển khai với ba mô hình khác nhau: (i) MLP với kiến trúc ẩn (128, 64, 32) kết hợp cơ chế dừng sớm để hạn chế quá khớp; (ii) Random Forest [20] với 200 cây, max_depth=20, class_weight= “balanced”; (iii) XGBoost [21] với 300 cây, max_depth=8, learning_rate=0,1. Bốn chiến lược kết hợp được khảo sát: Soft Voting, Weighted Voting (trọng số tỉ lệ F1-Macro), Confidence-based selection và Stacking với Logistic Regression làm *meta-learner* [15]. Trong Stacking, các vector xác suất dự đoán chéo được tạo bằng kỹ thuật kiểm định chéo 5-Fold trên tập huấn luyện sau SMOTE, sau đó nối thành meta-feature kích thước $3 \times C$ (C là số lớp); cách dùng dự đoán chéo giúp tránh rò rỉ thông tin từ các mô hình cơ sở sang *meta-learner*. Phương án tốt nhất được chọn theo hàm đánh giá điểm tổng hợp: $Score = 0,5 \times F1 - W + 0,3 \times F1 - M + 0,2 \times F1 - Legit$ nhằm cân bằng hiệu năng tổng thể với cải thiện lớp thiểu số. Gọi $P = (P_1, \dots, P_C) \in \mathbb{R}_C$ là vector xác suất đầu ra cuối của Tầng 1 (với P_{normal} là xác suất của lớp *legitimate*), thì dự đoán cứng là $\hat{y} = argmax(p)$.

2.3. Hệ suy diễn mờ có chọn lọc

FIS chỉ được kích hoạt khi đồng thời thỏa mãn: $\hat{y} = legitimate \text{ AND } P_{normal} \leq \min(\theta, GUARD)$, trong đó θ là ngưỡng định tuyến quyết định mẫu nào được chuyển sang FIS và $GUARD$ là ngưỡng chặn trên cứng giúp FIS không can thiệp vào các mẫu có xác suất legitimate cao. Giá trị của hai ngưỡng này được chọn thông qua khảo sát trên lưới giá trị 4 chiều ($\theta \times \alpha \times GUARD \times CONF$) trên tập validation độc lập tách từ tập test 20% ban đầu và không trùng với tập holdout. Trên lưới khảo sát, cấu hình tối ưu thu được giá trị của $\theta = 0,6$ và $GUARD=0,65$ cho hiệu năng tốt nhất trên tập kiểm thử về cả F1-Weighted và Legitimate-F1, đồng thời thỏa mãn ràng buộc $GUARD \geq \theta$ để phân tách rõ vai trò giữa hai ngưỡng. Bốn biến đầu vào FIS được tính như sau: (1) Normality Deviation $P_{dev} = \sqrt{1 - P_{normal}}$ đo độ lệch so với lớp legitimate; (2) Behavioral Abnormality $P_{abnorm} = \exp^{-\gamma \cdot D_{min}}$ với $D_{min} = \min_k \|x - \mu_k\|$ là khoảng cách Euclid tối thiểu đến centroid của các lớp tấn công μ_k trong nghiên cứu này giá trị $\gamma = 0,05$ được chọn dựa trên phân tích phân bố D_{min} trên tập Train. Giá trị này có ý nghĩa đối với các tập mờ *low/medium/high*. Nếu γ quá lớn P_{abnorm} sẽ bão hòa về 0 cho hầu hết mẫu, nếu γ quá nhỏ biến P_{abnorm} sẽ không phân biệt được; (3) Traffic Volume $V_{traffic} = clip(\frac{Header_Length}{10000}, 0, 1)$; và (4) Prediction Entropy $\hat{H} = -\frac{\sum_{c=1}^C p_c \log_2(p_c)}{\log_2(C)}$ đo độ bất định chuẩn hóa của p . Mỗi biến được mờ hóa bằng ba tập mờ tam giác *low/medium/high* trên $[0, 1]$; biến hệ quả biểu diễn mức độ ảnh hưởng dùng ba tập mờ tam giác với *low, medium, high* và một tập Gauss *critical* tập trung tại 1,0

($\sigma = 0,05$). Bốn biến được chọn để bù trừ các điểm yếu của Tầng 1: P_{normal} phản ánh độ tin cậy của Tầng 1 với lớp *legitimate*; P_{abnorm} đo khoảng cách đến centroid các lớp tấn công, xác định mẫu có gần vùng hành vi tấn công không; $V_{traffic}$ biểu thị quy mô lưu lượng; Prediction Entropy $\hat{H}$ khai thác chuẩn hóa của phân phối xác suất ensemble để đánh giá mức độ phân vân (bất định) của Tầng 1. Hệ luật Mamdani gồm 9 luật gồm: R1: nếu P_{dev} cao hoặc P_{abnorm} cao thì mức độ ảnh hưởng là “critical”; R2: nếu P_{dev} cao và $V_{traffic}$ cao thì mức độ ảnh hưởng là “critical”; R3: nếu P_{dev} và P_{abnorm} cùng “medium” thì mức độ ảnh hưởng là “high”; R4: nếu P_{dev} hoặc P_{abnorm} ở medium thì mức độ ảnh hưởng là “medium”; R5: nếu P_{dev} thấp thì mức độ ảnh hưởng thấp; R6: nếu $\hat{H}$ cao và P_{dev} medium thì mức độ ảnh hưởng là “high”; R7: nếu $\hat{H}$ cao và P_{abnorm} cao thì mức độ ảnh hưởng là “critical”; R8: nếu $\hat{H}$ thấp và P_{dev} cao thì mức độ ảnh hưởng là “critical”; R9 (bảo vệ legitimate): nếu $\hat{H}$ thấp và P_{dev} thấp thì mức độ ảnh hưởng thấp. Đầu ra giải mờ $S_{impact} \in [0,1]$ thu được bằng phương pháp trọng tâm. Trong đó: các luật R1, R2 tạo cảnh báo *critical* khi mẫu có độ lệch khỏi *legitimate* cao, gần vùng tấn công hoặc có lưu lượng bất thường; R3, R4 áp dụng cho các trường hợp trung bình với cảnh báo nghi ngờ chưa đủ mạnh để override quyết liệt; R5 và R9 có vai trò bảo vệ *legitimate* nếu mẫu gần vùng *legitimate* và mô hình đủ tự tin thì mức độ ảnh hưởng thấp tránh override sai; R6, R7, R8 bổ sung vai trò của Prediction Entropy – khi Tầng 1 phân vân hoặc mẫu lệch khỏi *legitimate* rõ rệt, FIS tăng mức cảnh báo để phục hồi các mẫu tấn công bị nhầm.

3. Kết quả và thảo luận

3.1. Cài đặt thực nghiệm

Toàn bộ thực nghiệm được triển khai, đánh giá trên bộ dữ liệu MQTTEEB-D [4] đây là bộ dữ liệu thu thập từ môi trường IoT thực gồm 6 lớp: *legitimate*, *bruteforce*, *DoS*, *flood*, *malformed*, *slowite*. Tổng số 222.813 mẫu được chia theo tỉ lệ 60/20/20 thành Train/Test/Holdout. Trên tập holdout có 44.514 mẫu trên 6 lớp được dùng để đánh giá cuối, phân phối nhãn: *legitimate* 7,9%, *dos* 18,8%, *bruteforce* 12,3%, *slowite* 9,0%, *flood* 4,6%. Toàn bộ thực nghiệm được thực hiện trên Google Colab với Python 3.10, imbalanced-learn [24], scikit-learn [25], xgboost và scikit-fuzzy. Seed random_state=42 được cố định để bảo đảm tái lập.

Các chỉ số đánh giá gồm: F1-Weighted và F1-Macro phản ánh hiệu năng tổng thể; Recall và F1 riêng cho *legitimate* là thang đo trọng tâm cho lớp thiểu số; Recall riêng cho *bruteforce* thể hiện khả năng phát hiện loại tấn công khó nhất; Average Precision (AP) per-class và Macro AP đo hiệu năng trên thang xác suất liên tục theo [8].

3.2. Kết quả

Để chứng minh hiệu quả của phương pháp được đề xuất, chúng tôi đã tiến hành phân tích các chỉ số phản ánh các khía cạnh khác nhau của một hệ IDS trên dữ liệu mất cân bằng gồm F1-Weighted (F1-W), F1-Macro (F1-Mac), Legitimate-Precision (Legit-P), Legitimate-Recall (Legit-R) và Bruteforce-Recall (Brute-R). Trong đó F1-W đo hiệu năng tổng thể có trọng số theo kích thước lớp, phản ánh chất lượng phân loại trên đa số lưu lượng; F1-Mac trung bình F1 không trọng số, nhạy với lớp thiểu số và bộc lộ tình trạng thiên về lớp đa số khi thấp hơn F1-W rõ rệt; Legit-R đo tỉ lệ phát hiện đúng lưu lượng hợp lệ, gắn trực tiếp với tỉ lệ cảnh báo giả trong vận hành thực; Legitimate-F1 (Legit-F1) cân bằng giữa precision và recall cho lớp thiểu số, thấy được sự đánh đổi (trade-off) giữa hai loại lỗi quan trọng nhất; Bruteforce-Recall (Brute-R) đo trực tiếp khả năng phát hiện loại tấn công khó nhất trong MQTTEEB-D, là chỉ số phản ánh đóng góp của Tầng 2 FIS một cách rõ ràng nhất. Bảng 1 trình bày so sánh hiệu năng giữa MLP, Ensemble Stacking và Hybrid Ensemble-FIS đề xuất được đánh giá trên cùng tập holdout. Trong đó precision của lớp *legitimate* (Legit-P) tăng 0,7 từ 0,6649 lên 0,7347 trong khi recall (Legit-R) giảm từ 0,7098 xuống 0,6482, nên Legit-F1 vẫn tăng nhẹ, đây là sự đánh đổi có chủ đích của tầng FIS.

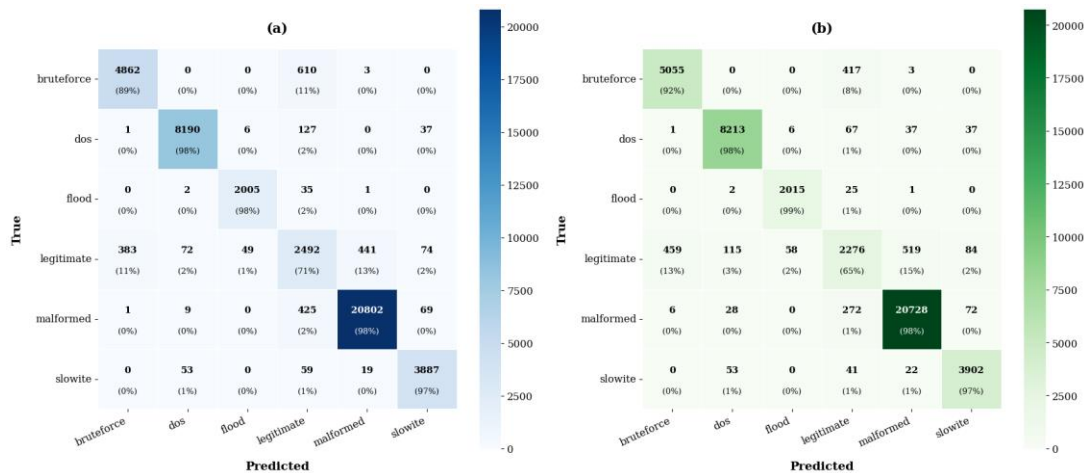
Bảng 1. So sánh hiệu năng của 3 mô hình trên tập holdout của MQTTEEB-D

Mô hình	F1-W	F1-Mac	Legit-P	Legit-R	Legit-F1	Brute-R
MLP	0,8625	0,8017	0,3974	0,2709	0,3222	0,9180
Ensemble Stacking	0,9450	0,9152	0,6649	0,7098	0,6866	0,8890
Hybrid Ensemble – FIS (đề xuất)	0,9465	0,9174	0,7347	0,6482	0,6888	0,9233

Để hiểu rõ tác động của FIS lên từng lớp tấn công, Bảng 2 trình bày recall mỗi lớp của Ensemble Baseline và Hybrid Ensemble-FIS trên tập holdout. Kết quả cho thấy lớp *bruteforce* được cải thiện rõ rệt với mức tăng 3% đây cũng là mục tiêu thiết kế của Tầng 2 FIS. Với bốn lớp tấn công còn lại kết quả gần như không thay đổi cho thấy FIS không gây nhiễu với các quyết định đã chắc chắn ở Tầng 1.

Bảng 2. So sánh recall per-class của Ensemble baseline và Hybrid Ensemble-FIS.
Số trong ngoặc là số mẫu phân loại đúng

Lớp	Recall Baseline	Recall Hybrid	Tỷ lệ tăng (%)	Nhận xét
Bruteforce	89% (4862)	92% (5055)	+3	Cải thiện tốt
Dos	98% (8190)	98% (8213)	≈ 0	Ổn định
Flood	98% (2005)	99% (2015)	+1	Cải thiện nhẹ
Legitimate	71% (2492)	65% (2276)	-6	Đánh đổi
Malformed	98% (20602)	98% (20728)	≈ 0	Ổn định
Slowite	97% (3887)	97% (3902)	≈ 0	Ổn định

**Hình 2.** Ma trận nhầm lẫn của 2 mô hình: Ensemble baseline (a) và Hybrid Ensemble-FIS (b)

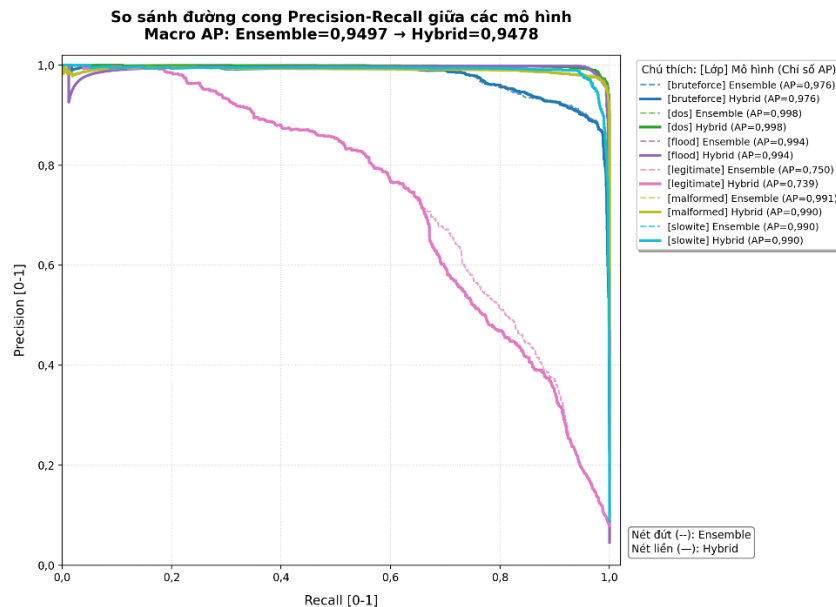
Dựa vào kết quả của Hình 2 có thể thấy mô hình Hybrid Ensemble-FIS phục hồi thêm 193 mẫu *bruteforce*, đưa recall lớp này từ 89% lên 92%. Recall của bốn lớp tấn công còn lại không thay đổi đáng kể. Đánh đổi xảy ra ở lớp *legitimate*: recall giảm 6 điểm (từ 71% xuống 65%), tương đương 216 mẫu *legitimate* bị chuyển sang attack, phân tán chủ yếu sang *malformed* (+78) và *bruteforce* (+76). Đây là sự đánh đổi có thể chấp nhận được trong bối cảnh IDS vì việc bỏ sót bruteforce thường có giá trị an ninh cao hơn việc xử lý cảnh báo giả từ lưu lượng hợp lệ. Tiếp theo để đánh giá mô hình Ensemble cơ sở và Hybrid Ensemble-FIS trên các lớp phân loại của tập Holdout qua chỉ số Average Precision, kết quả được trình bày ở Bảng 3.

Kết hợp Bảng 3 và Hình 3 có thể thấy các kết quả như sau: Thứ nhất, bốn lớp tấn công có chữ ký giao thức rõ rệt (dos, flood, malformed, slowite) đều có $AP \geq 0,990$ với hai đường baseline và hybrid gần như chùng khít, chứng minh rằng FIS không can thiệp vào các quyết định đã chắc chắn. Thứ hai, đường cong bruteforce ($AP \approx 0,976$) nằm ở mức trung-cao, duy trì precision $\geq 0,95$ trong phần lớn dải recall và chỉ tụt nhanh khi recall $> 0,9$ đây là vùng các mẫu bruteforce

khó nhận diện do hành vi mẫu tin giống đăng nhập hợp lệ. Thứ ba, đường legitimate (baseline 0,750; hybrid 0,739) khác biệt rõ rệt so với năm lớp còn lại, với khoảng cách lớn nhất giữa hai đường nằm ở vùng recall trung bình 0,5–0,8 chính là vùng FIS tác động mạnh nhất qua cơ chế blending. Thứ tư, độ chính xác trung bình Macro AP giảm nhẹ (-0,0019) trong đó lớp legitimate (-0,0114), năm lớp còn lại có độ chênh lệch $\leq 0,0002$ nằm trong sai số đo. Điều này xác nhận FIS chỉ tác động đúng vùng được thiết kế (ranh giới legitimate $\leftrightarrow$ brute force) mà không gây nhiễu cho các lớp khác.

Bảng 3. Độ chính xác trung bình (Average Precision (AP)) theo từng lớp

Lớp	AP Baseline	AP Hybrid	Chênh lệch AP
Brute force	0,9757	0,9759	+0,0002
Dos	0,9980	0,9980	-0,00002
Flood	0,9935	0,9935	+0,00001
legitimate	0,7501	0,7387	-0,0114
malformed	0,9907	0,9905	-0,0002
Slowite	0,9904	0,9904	+0,00003
Macro AP	0,9497	0,9478	-0,0019



Hình 3. Đường cong Precision-Recall cho 6 lớp

3.3. Thảo luận

Từ các kết quả thu được sau thực nghiệm của nghiên cứu này so với các nghiên cứu đã công bố trên MQTTEEB-D được trình bày ở Bảng 4. Nghiên cứu [26] cho kết quả F1-Mac = 0,99 với phương pháp Bagging và 0,98 cho CART trên bài toán phân loại 6 lớp, tuy nhiên kết quả này dựa trên kiểm chứng 80/20 đơn giản không có tập Validation độc lập, do đó có thể tồn tại vấn đề rò rỉ dữ liệu. Phương pháp đề xuất đạt F1-Mac = 0,9174 thấp hơn nhưng được kiểm chứng trên quy trình tách 60/20/20, trong đó 60% là tập Train, tập Test 20% tách tiếp ra 10% Test và 10% Validation và 20% là tập Holdout gồm 44.514 mẫu được tách riêng để đảm bảo không bị rò rỉ dữ liệu. Nghiên cứu [27] cũng là bài toán đa lớp, đạt F1-Mac = 0,86 với phương pháp Decision Tree (DT) và 0,85 với phương pháp GRU thấp hơn phương pháp đề xuất. Nghiên cứu [28] đạt F1-Mac = 0,96 nhưng trên bài toán nhị phân, về bản chất dễ hơn bài toán 6 lớp và không so sánh trực tiếp được. Đóng góp khác biệt của phương pháp đề xuất không chỉ đánh giá dựa trên F1-Mac mà còn phân tích per-class chi tiết (Brute-R = 0,9233, Legit-F1 = 0,6888) mà các công trình đã nêu

không công bố, ngoài ra còn có cơ chế hiệu chỉnh có chọn lọc và giải thích được thông qua FIS, đây cũng là hai khía cạnh có ý nghĩa thực tiễn cao cho triển khai an ninh mạng IoT.

Bảng 4. So sánh định lượng với các công trình trên cùng tập dữ liệu MQTTEEB-D

Công trình	Phương pháp	Bài toán	Phân chia dữ liệu	Acc	F1-Mac
[26]	Bagging	6 lớp	80/20	0,99	0,99
	CART (DT)			0,99	0,98
	1D CNN			0,641	0,59
[27]	DT	6 lớp	80/20	0,87	0,86
	GRU			0,86	0,85
[28]	DT	2 lớp	80/20	0,98	0,96
Đề xuất	Hybrid Ensemble-FIS	6 lớp	60/20/20	~0,95	0,917

Các công trình [26]-[28] chủ yếu tập trung vào mô hình phát hiện tổng quát và báo cáo hiệu năng chung, trong khi phương pháp đề xuất của chúng tôi không chỉ tối ưu hiệu năng phân loại tổng thể mà còn đi sâu vào vấn đề nhầm lẫn giữa *legitimate* và các lớp tấn công, đặc biệt là *bruteforce*. Bằng cách kết hợp Ensemble Stacking với FIS, phương pháp đề xuất duy trì F1-W và F1-Mac ở mức cao, đồng thời cải thiện Brute-R và giải thích được sự thay đổi của Legit-F1 thông qua Legit-P. Do đó, đóng góp chính của phương pháp không chỉ cải thiện hiệu năng mà còn là cơ chế hiệu chỉnh có điều kiện, có khả năng giải thích cho các lỗi phân loại khó trên MQTTEEB-D.

Kết quả thực nghiệm khẳng định rằng phần lớn cải thiện đến từ việc nâng cấp Tầng 1 từ MLP đơn sang Ensemble Stacking, với composite score tăng +0,1003 và Legit-F1 tăng gấp đôi. Điều này phù hợp với cơ sở lý thuyết về Ensemble: việc kết hợp ba mô hình có bias khác nhau (phi tuyến mạnh từ MLP, rule-based từ RF, gradient boosting từ XGBoost) cho phép meta-learner khai thác đóng góp đặc trưng của mỗi mô hình trên mỗi vùng không gian đặc trưng [13]. FIS đặt sau Ensemble đem lại cải thiện nhỏ về số tuyệt đối nhưng quan trọng về định tính: việc phục hồi 193 mẫu bruteforce dựa trên chỉ số đánh giá theo ngưỡng điểm có giá trị thực tế cao vì đây là loại tấn công nguy hiểm thường dẫn đến tổn hại tài khoản và xâm nhập sâu hơn. Điều này chứng minh Hybrid Ensemble-FIS có tiềm năng dùng để xếp hạng cảnh báo theo mức độ rủi ro đảm bảo được tính minh bạch của hệ thống IDS.

4. Kết luận

Kiến trúc hai tầng kết hợp Ensemble Stacking và Hệ suy diễn mờ có chọn lọc cho thấy hiệu quả phân vai rõ rệt: Tầng 1 đảm nhận phân loại tổng quát và đóng góp phần lớn cải thiện hiệu năng, Tầng 2 FIS đóng vai trò tinh chỉnh có chọn lọc ở ranh giới mờ giữa lớp *legitimate* và lớp *bruteforce*, vốn là điểm khó phân biệt nhất trong dữ liệu MQTT khó phân, nâng recall của lớp tấn công bruteforce từ 89% lên 92%. Kết quả phân tích trên nhiều chỉ số cho thấy phương pháp đề xuất cải thiện ở thang quyết định cứng (F1, Recall) nhưng lại không cải thiện ở thang xác suất liên tục (AP). Cụ thể, Hybrid Ensemble-FIS làm Macro AP giảm nhẹ 0,0019 điểm, sự giảm nhẹ này phần lớn đến từ lớp *legitimate*. Đây là cơ sở để đánh giá có ý nghĩa thực tiễn khi đánh giá các hệ IDS trên dữ liệu mất cân bằng. Các hướng nghiên cứu tiếp theo bao gồm: tinh chỉnh cơ chế kết hợp xác suất để khôi phục AP của lớp *legitimate*; mở rộng FIS sang các nhánh tấn công khác; tích hợp luật mờ học được hướng tới hệ Neuro-Fuzzy IDS hoàn chỉnh; và kiểm chứng tính tổng quát trên các bộ dữ liệu MQTT khác như MQTT-IoT-IDS2020 và MQTTset. Về khía cạnh triển khai, hệ thống đề xuất phù hợp với gateway IoT vận hành với ngưỡng quyết định cố định, đặc biệt trong các kịch bản đòi hỏi cả hiệu năng phát hiện cao và khả năng giải thích quyết định.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Giao thông vận tải (ĐH GTVT) trong đề tài mã số T2026-TTKHCN-001.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] OASIS, "MQTT Version 5.0 OASIS Standard," OASIS Open, 7 March 2019. [Online]. Available: <https://docs.oasis-open.org/mqtt/mqtt/v5.0/mqtt-v5.0.html>. [Accessed May 24, 2026].
- [2] ISO/IEC, "Information technology — Message Queuing Telemetry Transport (MQTT) v3.1.1," International Standard ISO/IEC 20922:2016, 2016.
- [3] Y. Kombate, P. Houngue, and O. Samuel, "Securing MQTT: Unveiling vulnerabilities and innovating cyber range solutions," *Procedia Computer Science*, vol. 241, pp. 69-76, 2024, doi: 10.1016/j.procs.2024.08.012.
- [4] A. Aqachtoul *et al.*, "MQTTEEB-D: A real-world IoT cybersecurity dataset for AI-powered threat detection in MQTT networks," *Data in Brief*, vol. 62, 2025, doi: 10.1016/j.dib.2025.111897.
- [5] I. Vaccari, M. Aiello, and E. Cambiaso, "SlowITe, a novel denial of service attack affecting MQTT," *Sensors*, vol. 20, no. 10, 2020, doi: 10.3390/s20102932.
- [6] A. Abdelkhalek and M. Mashaly, "Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning," *The Journal of Supercomputing*, vol. 79, no. 10, pp. 10611-10641, 2023, doi: 10.1007/s11227-023-05073-x.
- [7] H. Ren, Y. Tang, W. Dong, S. Ren, and L. Jiang, "DUEN: Dynamic ensemble handling class imbalance in network intrusion detection," *Expert Systems with Applications*, vol. 229, 2023, doi: 10.1016/j.eswa.2023.120420.
- [8] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0118432.
- [9] L. Fischer and P. Wollstadt, "Precision and recall reject curves for classification," *arXiv preprint arXiv:2308.08381*, 2023, doi: 10.48550/arXiv.2308.08381.
- [10] A. M. Banaamah and I. Ahmad, "Intrusion detection in IoT using deep learning," *Sensors*, vol. 22, no. 21, 2022, doi: 10.3390/s22218417.
- [11] P. Sinha, D. Sahu, S. Prakash, T. Yang, R. S. Rathore, and V. K. Pandey, "A high performance hybrid LSTM CNN secure architecture for IoT environments using deep learning," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-94500-5.
- [12] H. Hindy, E. Bayne, M. Bures, R. Atkinson, C. Tachtatzis, and X. Bellekens, "Machine learning based IoT intrusion detection system: An MQTT case study (MQTT-IoT-IDS2020 dataset)," *arXiv preprint arXiv:2006.15340v3*, Nov. 2020.
- [13] M. A. Khan, M. A. Khan, S. U. Jan, J. Ahmad, S. S. Jamal, A. A. Shah, N. Pitropakis, and W. J. Buchanan, "A deep learning-based intrusion detection system for MQTT enabled IoT," *Sensors*, vol. 21, no. 21, 2021, doi: 10.3390/s21217016.
- [14] A. Almotairi, S. Atawneh, O. A. Khashan, and N. M. Khafajah, "Enhancing intrusion detection in IoT networks using machine learning-based feature selection and ensemble models," *Systems Science & Control Engineering*, vol. 12, no. 1, 2024, doi: 10.1080/21642583.2024.2321381.
- [15] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992, doi: 10.1016/S0893-6080(05)80023-1.
- [16] X. Qiu, L. Shi, and P. Fan, "A cooperative intrusion detection system for internet of things using fuzzy logic and ensemble of convolutional neural networks," *Scientific Reports*, vol. 15, 2025, doi: 10.1038/s41598-025-99938-1.
- [17] M. Almseidin, J. Al-Sawwa, M. Alkasassbeh, and M. Alweshah, "On detecting distributed denial of service attacks using fuzzy inference system," *Cluster Computing*, vol. 26, no. 2, pp. 1337-1351, 2023, doi: 10.1007/s10586-022-03657-5.
- [18] S. S. Pandi, R. D. R. Ramani, A. Senthilselvi, S. Dhanasekaran, B. Kalpana, and N. A. Balaji, "Advancing IoT security with FLAME: A hybrid approach combining fuzzy logic and artificial lizard search optimization," *Computers & Security*, vol. 145, 2024, doi: 10.1016/j.cose.2024.103984.
- [19] A. H. Abu Saq, A. Zainal, B. A. S. Al-Rimy, A. Alyami, and H. A. Abosag, "Intrusion detection in IoT using Gaussian fuzzy mutual information-based feature selection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 17564-17571, 2024, doi: 10.48084/etasr.8268.
- [20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.

-
- [21] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
- [22] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," *IEEE Transactions on Computers*, vol. C-26, no. 12, pp. 1182-1191, 1977, doi: 10.1109/TC.1977.1674779.
- [23] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338-353, 1965, doi: 10.1016/S0019-9958(65)90241-X.
- [24] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1-5, 2017, doi: 10.5555/3122009.3122026.
- [25] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, 2011, doi: 10.5555/1953048.2078195.
- [26] H. Allaga, M. Biniz, and A. Farchane, "MQTTEEB-D: A high-fidelity benchmark for real-time MQTT anomaly detection using machine learning techniques," *Ad Hoc Networks*, vol. 181, 2026, Art. no. 104062, doi: 10.1016/j.adhoc.2025.104062.
- [27] K. Karam, A. Aqachtoul, A. Elamrani, M. Najib, N. Rafalia, I. Moumen, J. Abouchabaka, and M. Bakhouya, "ISAAF: an IoT security and attack prevention framework using AI-driven predictive analytics," *Scientific Reports*, vol. 15, 2025, Art. no. 44913, doi: 10.1038/s41598-025-28516-2.
- [28] A. K. Alhossainy and M. Malekzadeh, "AI Powered Threat Detection Framework for Security Enhancement of MQTT based IoT Networks," *International Transactions on Electrical Engineering and Computer Science*, vol. 5, no. 1, 2026, doi: 10.62760/iteecs.5.1.2026.171.