TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT TP. HỒ CHÍ MINH

KHOA ĐIỆN - ĐIỆN TỬ

BỘ MÔN KỸ THUẬT MÁY TÍNH

ĐỒ ÁN TỐT NGHIỆP

ĐỀ TÀI:

THIẾT KẾ VÀ ĐÁNH GIÁ HIỆU QUẢ TĂNG CƯỜNG CỦA HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG ỨNG DỤNG MÔ HÌNH KHUẾCH TÁN TRONG AN TOÀN MẠNG IOMT

Ngành: Công nghệ Kỹ thuật Máy tính

Sinh viên thực hiện:

Bùi Nguyễn Bảo Châu - MSSV: 22119168

Hồ Nguyễn Minh Vĩ - MSSV: 22119256

Giảng viên hướng dẫn: PGS.TS. Võ Minh Huân

TP. Hồ Chí Minh - Tháng 6/2026

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ KỸ THUẬT TP. HỒ CHÍ MINH

KHOA ĐIỆN - ĐIỆN TỬ

BỘ MÔN KỸ THUẬT MÁY TÍNH

ĐỒ ÁN TỐT NGHIỆP

ĐỀ TÀI:

THIẾT KẾ VÀ ĐÁNH GIÁ HIỆU QUẢ TĂNG CƯỜNG CỦA HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG ỨNG DỤNG MÔ HÌNH KHUẾCH TÁN TRONG AN TOÀN MẠNG IOMT

Ngành: Công nghệ Kỹ thuật Máy tính

Sinh viên thực hiện:

Bùi Nguyễn Bảo Châu - MSSV: 22119168

Hồ Nguyễn Minh Vĩ - MSSV: 22119256

Giảng viên hướng dẫn: PGS.TS. Võ Minh Huân

TP. Hồ Chí Minh - Tháng 6/2026

BẢN NHẬN XÉT KHÓA LUẬN TỐT NGHIỆP

(Dành cho giảng viên hướng dẫn)

Đề tài: THIẾT KẾ VÀ ĐÁNH GIÁ HIỆU QUẢ TĂNG CƯỜNG CỦA HỆ THỐNG PHÁT HIỆN XÂM NHẬP MẠNG ỨNG DỤNG MÔ HÌNH KHUẾCH TÁN TRONG AN TOÀN MẠNG IOMT

Sinh viên: + ……Bùi Nguyễn Bảo Châu……. MSSV: 22119168

+ ……Hồ Nguyên Minh Vĩ…….. MSSV: 22119256 Hướng dẫn: ……..…PGS. TS Võ Minh Huân……..…….

Nhận xét bao gồm các nội dung sau đây:

1. *Tính hợp lý trong cách đặt vấn đề và giải quyết vấn đề; ý nghĩa khoa học và thực tiễn:*

   Đặt vấn đề rõ ràng, mục tiêu cụ thể; đề tài có tính mới, cấp thiết; đề tài có khả năng ứng dụng, tính sáng tạo. Đề tài có tính cấp thiết, mục tiêu rõ ràng, hướng nghiên cứu phù hợp và có ý nghĩa ứng dụng trong lĩnh vực an toàn mạng IoMT.

2. *Phương pháp thực hiện/ phân tích/ thiết kế: Phương pháp hợp lý và tin cậy dựa trên cơ sở lý thuyết; có phân tích và đánh giá phù hợp; có tính mới và tính sáng tạo.*

   Phương pháp nghiên cứu hợp lý, dựa trên cơ sở lý thuyết và thực nghiệm, quy trình triển khai rõ ràng, có tính sáng tạo.

3. *Kết quả thực hiện/ phân tích và đánh giá kết quả/ kiểm định thiết kế:*

   Phù hợp với mục tiêu đề tài; phân tích và đánh giá / kiểm thử thiết kế hợp lý; có tính sáng tạo/ kiểm định chặt chẽ và đảm bảo độ tin cậy.

   Kết quả đạt được phù hợp với mục tiêu đề ra, được phân tích và đánh giá đầy đủ, có độ tin cậy.

4. *Kết luận và đề xuất:*

   Kết luận phù hợp với nội dung nghiên cứu, các đề xuất có tính khả thi và định hướng phát triển tốt.

5. *Hình thức trình bày và bố cục báo cáo:*

   Văn phong nhất quán, bố cục hợp lý, cấu trúc rõ ràng, đúng định dạng mẫu; có tính hấp dẫn, thể hiện năng lực tốt, văn bản trau chuốt.

   Báo cáo trình bày rõ ràng, bố cục hợp lý, văn phong mạch lạc, đúng quy định.

6. *Kỹ năng chuyên nghiệp và tính sáng tạo:*

   Thể hiện các kỹ năng giao tiếp, kỹ năng làm việc nhóm, và các kỹ năng chuyên nghiệp khác trong việc thực hiện đề tài.

   Thể hiện tinh thần nghiên cứu nghiêm túc, khả năng làm việc nhóm và tư duy sáng tạo tốt.

7. *Tài liệu trích dẫn*

   Tính trung thực trong việc trích dẫn tài liệu tham khảo; tính phù hợp của các tài liệu trích dẫn; trích dẫn theo đúng chỉ dẫn APA.

Tài liệu tham khảo phù hợp, được trích dẫn đầy đủ và cơ bản tuân thủ theo chuẩn.

8. *Đánh giá về sự trùng lặp của đề tài*

   Cần khẳng định đề tài có trùng lặp hay không? Nếu có, đề nghị ghi rõ mức độ, tên đề tài, nơi công bố, năm công bố của đề tài đã công bố.

   Không phát hiện dấu hiệu trùng lặp đáng kể với các công trình đã công bố.

9. *Những nhược điểm và thiếu sót, những điểm cần được bổ sung và chỉnh sửa**

   Cần bổ sung thêm phân tích so sánh với nhiều phương pháp hiện đại và mở rộng phạm vi đánh giá thực nghiệm.

10. *Nhận xét tinh thần, thái độ học tập, nghiên cứu của sinh viên*

    Sinh viên có tinh thần học tập nghiêm túc, chủ động nghiên cứu và hoàn thành tốt các nội dung của đề tài.

Đề nghị của giảng viên hướng dẫn

Ghi rõ: "Báo cáo đạt/ không đạt yêu cầu của một khóa luận tốt nghiệp kỹ sư, và được phép/ không được phép bảo vệ khóa luận tốt nghiệp"

Báo cáo đạt yêu cầu của một khóa luận tốt nghiệp kỹ sư, và được phép được phép bảo vệ khóa luận tốt nghiệp.

Tp. HCM, ngày 27 tháng 6 năm 2026

Người nhận xét

Võ Minh Huân

LỜI CẢM ƠN

Để hoàn thành đề tài "Thiết kế và đánh giá hiệu quả tăng cường của hệ thống phát hiện xâm nhập mạng ứng dụng mô hình khuếch tán trong an toàn mạng IoMT", trước hết, chúng tôi xin chân thành cảm ơn PGS.TS. Võ Minh Huân — Giảng viên Khoa Điện - Điện tử — đã rất kiên nhẫn hướng dẫn, định hướng và chia sẻ cùng hỗ trợ chúng tôi trong suốt quá trình thực hiện đề tài.

Chúng tôi xin cảm ơn toàn thể quý thầy cô trong Khoa Điện - Điện tử, Trường Đại học Công nghệ Kỹ thuật TP. Hồ Chí Minh, cũng như bạn bè cùng khóa đã hợp tác và giúp đỡ nhóm trong quá trình thực hiện đồ án.

Ngoài ra, chúng tôi cũng xin cảm ơn gia đình và bạn bè đã luôn ủng hộ, động viên và tạo điều kiện tốt nhất để chúng tôi có thể hoàn thành đề tài trong thời gian qua.

Do kiến thức còn hạn chế nên trong quá trình thực hiện không tránh khỏi những thiếu sót. Chúng tôi rất mong nhận được nhiều ý kiến đóng góp và chỉ bảo của quý thầy cô để đề tài được hoàn thiện hơn.

Chúng tôi xin chân thành cảm ơn!

TP. Hồ Chí Minh, ngày ... tháng 6 năm 2026

Nhóm sinh viên thực hiện

Bùi Nguyễn Bảo Châu **Hồ Nguyễn Minh Vĩ**

LỜI CAM ĐOAN

Chúng tôi xin cam đoan đề tài "Thiết kế và đánh giá hiệu quả tăng cường của hệ thống phát hiện xâm nhập mạng ứng dụng mô hình khuếch tán trong an toàn mạng IoMT" là công trình nghiên cứu của chính nhóm, được thực hiện dưới sự hướng dẫn của PGS.TS. Võ Minh Huân.

Toàn bộ số liệu và kết quả thực nghiệm trình bày trong báo cáo là trung thực, do nhóm tự triển khai và thu được từ quá trình huấn luyện — kiểm thử trên tập dữ liệu CICIoMT2024. Mọi thí nghiệm đều cố định hạt giống ngẫu nhiên (seed = 42) nên hoàn toàn có thể tái lập lại. Các đoạn mã nguồn, mô hình và quy trình xử lý do nhóm tự cài đặt bằng Python.

Mọi tham khảo từ tài liệu, công trình của tác giả khác đều được trích dẫn đầy đủ và rõ ràng trong phần Tài liệu tham khảo. Nếu phát hiện có bất kỳ sự gian lận nào, chúng tôi xin hoàn toàn chịu trách nhiệm trước Hội đồng cũng như kết quả đồ án của mình.

TP. Hồ Chí Minh, ngày ... tháng 6 năm 2026

Nhóm sinh viên thực hiện

Bùi Nguyễn Bảo Châu **Hồ Nguyễn Minh Vĩ**

TÓM TẮT ĐỀ TÀI

Sự bùng nổ của Internet of Medical Things (IoMT) trong những năm gần đây đã mang lại nhiều lợi ích to lớn cho ngành y tế, song cũng đặt ra những thách thức nghiêm trọng về an ninh mạng. Các thiết bị y tế kết nối như máy theo dõi bệnh nhân, bơm tiêm điện hay máy tạo nhịp tim thường bị hạn chế về tài nguyên phần cứng, khiến việc triển khai các giải pháp bảo mật truyền thống trở nên khó khả thi. Trong bối cảnh đó, hệ thống phát hiện xâm nhập (Intrusion Detection System — IDS) dựa trên học máy nổi lên như một hướng tiếp cận hiệu quả để bảo vệ hạ tầng mạng IoMT.

Tuy nhiên, một vấn đề cốt lõi và dai dẳng trong lĩnh vực này là tình trạng mất cân bằng dữ liệu nghiêm trọng: lưu lượng mạng bình thường chiếm phần lớn tuyệt đối trong khi các loại tấn công nguy hiểm lại hiếm gặp. Điều này khiến cho các mô hình học máy bị thiên vị, dự đoán tốt lớp đa số nhưng bỏ sót các mối đe dọa thực sự.

Đề tài này đề xuất và triển khai một pipeline hoàn chỉnh trên tập dữ liệu CICIoMT2024 — bộ dữ liệu an ninh mạng IoMT mới với 18 loại tấn công được chia thành 5 nhóm trên 40 thiết bị thực tế. Nhóm so sánh có hệ thống ba kịch bản tăng cường dữ liệu: Baseline (không tăng cường), SMOTE (nội suy tuyến tính) và TabDDPM (mô hình khuếch tán học sâu), trên các bộ phân loại thống nhất.

Kết quả thực nghiệm cho thấy: ở quy mô dữ liệu đầy đủ, các phương pháp đạt mức bão hòa và cho hiệu năng tương đương. Tuy nhiên, trong kịch bản nhãn khan hiếm (low-data) — sát với thực tế triển khai IoMT — mô hình khuếch tán TabDDPM thể hiện ưu thế vượt trội. Điều này khẳng định giá trị của dữ liệu tổng hợp chất lượng cao từ mô hình khuếch tán trong việc nâng cao khả năng phát hiện các lớp tấn công thiểu số — yếu tố then chốt quyết định hiệu quả thực tế của một hệ thống IDS trong môi trường mạng y tế.

LỜI MỞ ĐẦU

Cuộc cách mạng công nghiệp lần thứ tư đã đưa Internet vạn vật (IoT) thâm nhập sâu rộng vào mọi lĩnh vực của đời sống, trong đó y tế là một trong những ngành được hưởng lợi nhiều nhất. Sự ra đời của Internet vạn vật y tế (IoMT) cho phép theo dõi sức khỏe bệnh nhân liên tục, từ xa và theo thời gian thực, mở ra kỷ nguyên mới cho y học chính xác và chăm sóc sức khỏe chủ động. Tuy nhiên, đi kèm với những lợi ích to lớn là các nguy cơ an ninh mạng ngày càng phức tạp, đe dọa trực tiếp đến sự an toàn và quyền riêng tư của người bệnh.

Trong số các giải pháp bảo vệ hạ tầng IoMT, hệ thống phát hiện xâm nhập dựa trên học máy nổi lên như một hướng đi đầy triển vọng nhờ khả năng phát hiện cả những loại tấn công chưa từng biết. Thế nhưng, hiệu quả của các hệ thống này bị cản trở nghiêm trọng bởi vấn đề mất cân bằng dữ liệu — khi các mẫu tấn công nguy hiểm nhưng hiếm gặp bị lấn át bởi lưu lượng bình thường, khiến mô hình học máy bỏ sót đúng những mối đe dọa cần phát hiện nhất.

Xuất phát từ thực tiễn đó, đề tài "Thiết kế và đánh giá hiệu quả tăng cường của hệ thống phát hiện xâm nhập mạng ứng dụng mô hình khuếch tán trong an toàn mạng IoMT" được thực hiện nhằm nghiên cứu và đánh giá một cách có hệ thống vai trò của các kỹ thuật tăng cường dữ liệu — đặc biệt là mô hình khuếch tán TabDDPM — trong việc cải thiện khả năng phát hiện các lớp tấn công thiểu số. Đề tài không chỉ dừng lại ở việc so sánh trên dữ liệu đầy đủ, mà còn đi sâu vào kịch bản nhãn khan hiếm, vốn phản ánh đúng những khó khăn thực tế trong thu thập dữ liệu an ninh mạng y tế.

Báo cáo được trình bày thành năm chương, đi từ tổng quan và cơ sở lý thuyết, qua thiết kế và triển khai hệ thống, đến phân tích kết quả thực nghiệm và kết luận. Nhóm tác giả hy vọng những kết quả đạt được sẽ đóng góp một góc nhìn hữu ích cho hướng nghiên cứu ứng dụng mô hình sinh dữ liệu sâu vào bài toán an ninh mạng IoMT.

MỤC LỤC

LỜI CẢM ƠN 3
LỜI CAM ĐOAN 6
TÓM TẮT ĐỀ TÀI 7
LỜI MỞ ĐẦU 8
MỤC LỤC 1
DANH MỤC HÌNH VẼ 4
DANH MỤC BẢNG BIỂU 5
DANH MỤC CÁC TỪ VIẾT TẮT 6
CHƯƠNG 1: GIỚI THIỆU 7
1.1. Tổng quan 7
1.2. Mục tiêu đề tài 8
1.3. Đối tượng nghiên cứu 9
1.4. Phạm vi nghiên cứu 9
1.5. Bố cục đồ án 9
1.6. Phương pháp nghiên cứu 10
CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 11
2.1. IoMT và an ninh mạng 11
2.1.1. Định nghĩa và đặc điểm IoMT 11
2.1.2. Các loại tấn công phổ biến trong CICIoMT2024 12
2.1.3. Hệ thống phát hiện xâm nhập (IDS) 14
2.2. Mất cân bằng dữ liệu (Class Imbalance) 15
2.2.1. Định nghĩa và nguyên nhân 15
2.2.2. Tác hại đối với mô hình học máy 16
2.2.3. Các hướng giải quyết 16
2.3. Các phương pháp xử lý mất cân bằng được so sánh 17

2.3.1. Baseline (Không tăng cường) 17
2.3.2. SMOTE — Synthetic Minority Over-sampling Technique 17
2.3.3. TabDDPM — Denoising Diffusion Probabilistic Model 18
2.3.4. Tổng quan và so sánh các mô hình tạo sinh 21
2.4. Các chỉ số (Metrics) đánh giá 21
2.4.1. Các đại lượng cơ bản từ ma trận nhầm lẫn 21
2.4.2. Precision, Recall và F1-score 22
2.4.3. Macro F1-score — Chỉ số cốt lõi 22
2.4.4. Ví dụ minh họa sự khác biệt Macro và Weighted F1 23
CHƯƠNG 3: THIẾT KẾ VÀ TRIỂN KHAI 24
3.1. Tổng quan kiến trúc hệ thống 24
3.2. Quy trình tiền xử lý dữ liệu 26
3.2.1. Tải dữ liệu và tham số cơ bản 26
3.2.2. Ánh xạ nhãn (CLASS_MAP) 26
3.2.3. Lấy mẫu phân tầng (Stratified Sampling) 26
3.2.4. Làm sạch và chuẩn hóa 27
3.2.5. Thống kê phân bố dữ liệu sau tiền xử lý 29
3.3. Kiến trúc đường ống tăng cường dữ liệu 32
3.3.1. Kịch bản Baseline (Đường cơ sở) 32
3.3.2. Phương pháp nội suy tuyến tính (SMOTE) 33
3.3.3. Mô hình khuếch tán dạng bảng (TabDDPM) 33
3.4. Quy trình huấn luyện và cấu hình bộ phân loại 35
3.5. Cơ chế kiểm soát chất lượng mẫu sinh 36
3.5.1. Tầng 1 — Lọc theo độ lệch phân phối MMD 36
3.5.2. Tầng 2 — Lọc dị thường bằng IsolationForest 36
CHƯƠNG 4: KẾT QUẢ VÀ ĐÁNH GIÁ 38
4.1. Môi trường triển khai thực nghiệm 38

4.2. Thí nghiệm khan hiếm dữ liệu — ưu thế của TabDDPM 38
4.2.1. Thiết kế thí nghiệm.....38
4.3.2. Kết quả so sánh theo từng nhóm 39
4.3.3. Phân tích lớp thiểu số Spoofing.....40
4.3.4. Chất lượng mẫu sinh của TabDDPM 41
4.3.5. Phân tích ma trận nhầm lẫn 42
4.3.6. Trực quan hóa phân bố mẫu sinh 43
4.4. Hạn chế của nghiên cứu.....45
4.5. Nhận xét tổng hợp 45
CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 47
5.1. Kết luận.....47
5.2. Hướng phát triển trong tương lai 48
TÀI LIỆU THAM KHẢO 50

DANH MỤC HÌNH VẼ

Hình 2.1. Hệ sinh thái Internet of Medical Things (IoMT).....11

Hình 2.2. Mô hình chung của hệ thống IDS dựa trên học máy.....14

Hình 3.1. Sơ đồ pipeline thực nghiệm tổng thể.....25

Hình 3.2. Phân bố số mẫu các nhóm (thang logarit) minh họa mức độ mất cân bằng..30

Hình 3.3. Top 15 đặc trưng quan trọng nhất theo RandomForest.....31

Hình 3.4. Phân bố sáu nhóm trong không gian đặc trưng (chiếu t-SNE 2D).....32

Hình 3.5. Lưu đồ tăng cường dữ liệu TabDDPM theo từng nhóm 34

Hình 4.1. Kết quả chạy dataset CIC-IDS2017 với model phân loại ANN.....40

Hình 4.2. F1-score của lớp Spoofing — TabDDPM vượt trội trong điều kiện chỉ 200 mẫu thật 41

Hình 4.3. Độ lệch phân phối MMD của mẫu sinh — đều thấp hơn nhiều so với ngưỡng loại bỏ 42

Hình 4.4. Số mẫu mỗi nhóm trước và sau tăng cường (kịch bản khan hiếm).....44

Hình 4.5. Phân bố mẫu thật và mẫu sinh TabDDPM của lớp Spoofing (chiếu PCA 2D) 45

DANH MỤC BẢNG BIỂU

Bảng 1.1. Sáu nhóm phân loại trong phạm vi nghiên cứu 9

Bảng 1.2. Cấu trúc nội dung báo cáo.....10

Bảng 2.1. Siêu tham số SMOTE 18

Bảng 2.2. Siêu tham số TabDDPM 20

Bảng 2.3. Ví dụ minh họa Macro F1 và Weighted F1 23

Bảng 3.1. Bảng ánh xạ CLASS_MAP từ CICIoMT2024 26

Bảng 3.2. Chiến lược lấy mẫu phân tầng trên CICIoMT2024 27

Bảng 3.3. Phân nhóm các đặc trưng đầu vào 28

Bảng 3.4. Danh sách đầy đủ 45 đặc trưng đầu vào 29

Bảng 3.5. Phân bố số mẫu các nhóm trong tập huấn luyện.....30

Bảng 3.6. Kế hoạch tăng cường dữ liệu per-nhóm (TabDDPM) 35

Bảng 3.7. Tổng hợp cấu hình ba kịch bản thực nghiệm.....36

Bảng 4.1. F1-score theo nhóm — kịch bản khan hiếm dữ liệu (bộ phân loại RF).....39

Bảng 4.2. Chỉ số chất lượng mẫu sinh TabDDPM theo nhóm.....41

Bảng 4.3. Ma trận nhầm lẫn của tổ hợp TabDDPM + RandomForest (kịch bản khan hiếm).....42

Bảng 4.4. Ma trận nhầm lẫn của tổ hợp Baseline + RandomForest (kịch bản khan hiếm).....43

Bảng 4.5. Ma trận nhầm lẫn của tổ hợp SMOTE + RandomForest (kịch bản khan hiếm).....43

DANH MỤC CÁC TỪ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
IoMT	Internet of Medical Things	Internet vạn vật y tế
IoT	Internet of Things	Internet vạn vật
IDS	Intrusion Detection System	Hệ thống phát hiện xâm nhập
ML	Machine Learning	Học máy
DDPM	Denoising Diffusion Probabilistic Model	Mô hình khuếch tán khử nhiễu xác suất
TabDDPM	Tabular DDPM	Mô hình khuếch tán cho dữ liệu dạng bảng
SMOTE	Synthetic Minority Over-sampling Technique	Kỹ thuật sinh mẫu thiểu số tổng hợp
RF	Random Forest	Rừng ngẫu nhiên
MMD	Maximum Mean Discrepancy	Độ lệch trung bình cực đại
DDoS	Distributed Denial of Service	Từ chối dịch vụ phân tán
DoS	Denial of Service	Từ chối dịch vụ
MQTT	Message Queuing Telemetry Transport	Giao thức truyền thông điệp nhẹ
CoAP	Constrained Application Protocol	Giao thức ứng dụng ràng buộc
ARP	Address Resolution Protocol	Giao thức phân giải địa chỉ
F1	F1-score	Trung bình điều hòa Precision và Recall
TP / FP	True / False Positive	Dương tính thật / Dương tính giả
TN / FN	True / False Negative	Âm tính thật / Âm tính giả
LR	Learning Rate	Tốc độ học
CPU / GPU	Central / Graphics Processing Unit	Bộ xử lý trung tâm / đồ họa

CHƯƠNG 1: GIỚI THIỆU

1.1. Tổng quan

Internet of Medical Things (IoMT) là hệ sinh thái các thiết bị y tế được kết nối qua Internet, bao gồm máy theo dõi nhịp tim, máy đo huyết áp tự động, bơm tiêm điện thông minh và các cảm biến sinh học [2]. Số lượng thiết bị IoMT toàn cầu tăng trưởng vượt bậc, hứa hẹn mang lại đột phá trong chăm sóc sức khỏe từ xa, nhưng kéo theo những thách thức an ninh mạng nghiêm trọng [2, 3].

Các thiết bị IoMT lưu trữ dữ liệu y tế cực kỳ nhạy cảm song lại bị hạn chế về tài nguyên phần cứng, không đủ năng lực vận hành tường lửa hay phần mềm diệt virus thế hệ mới [3]. Bên cạnh đó, môi trường IoMT ưu tiên các giao thức nhẹ như MQTT và CoAP, vốn thiếu cơ chế mã hóa bảo mật sâu [4]. Các hình thức tấn công nhắm vào IoMT ngày càng đa dạng: từ DDoS/DoS làm tê liệt hạ tầng giám sát bệnh nhân, ARP Spoofing đánh chặn dữ liệu y tế, đến khai thác giao thức MQTT qua Malformed Data và Connect/Publish Flood [4].

Trước thực trạng đó, Học máy (Machine Learning — ML) nổi lên như hướng tiếp cận tối ưu cho Hệ thống phát hiện xâm nhập (IDS). Khác với IDS dựa trên dấu hiệu (signature-based) vốn chỉ phát hiện được các tấn công đã biết, mô hình ML có khả năng tổng quát hóa sang các dạng tấn công chưa từng thấy (biến thể, zero-day) [5, 7]. Tuy nhiên, thách thức cốt lõi của IDS dựa trên ML là tình trạng mất cân bằng dữ liệu nghiêm trọng. Sự mất cân bằng này khiến mô hình bị thiên vị về lớp đa số và gần như bỏ sót các tấn công thiểu số — vốn nguy hiểm hơn nhiều [6, 7].

Tập dữ liệu CICIoMT2024 được lựa chọn làm nền tảng thực nghiệm bởi đây là một trong những bộ dữ liệu an ninh mạng IoMT mới và toàn diện nhất hiện nay, được xây dựng từ 40 thiết bị y tế thực tế với đa dạng giao thức và hình thức tấn công. So với các bộ dữ liệu IoT tổng quát trước đây, CICIoMT2024 phản ánh chính xác hơn đặc thù của môi trường y tế, bao gồm cả các tấn công nhắm vào giao thức MQTT vốn phổ biến trong thiết bị y tế kết nối. Tính thời sự và độ tin cậy của bộ dữ liệu này là cơ sở quan trọng bảo đảm giá trị thực tiễn của các kết quả nghiên cứu.

Theo nhiều dự báo của ngành công nghiệp, số lượng thiết bị y tế kết nối sẽ tiếp tục tăng trưởng mạnh trong thập kỷ tới, kéo theo bề mặt tấn công (attack surface) ngày càng

mở rộng. Hậu quả của một cuộc tấn công thành công vào hạ tầng IoMT không chỉ dừng ở thiệt hại tài chính hay rò rỉ dữ liệu, mà còn có thể đe dọa trực tiếp đến tính mạng bệnh nhân — ví dụ, một cuộc tấn công DDoS làm tê liệt hệ thống giám sát nhịp tim trong phòng hồi sức tích cực có thể khiến cảnh báo nguy cấp không đến kịp với đội ngũ y tế. Chính vì mức độ nghiêm trọng đặc thù này, an ninh mạng IoMT đòi hỏi các giải pháp phát hiện xâm nhập có độ nhạy cao đối với cả những loại tấn công hiếm gặp nhất.

Đề tài này so sánh có hệ thống các phương pháp tăng cường dữ liệu thuộc hai trường phái khác nhau: SMOTE (nội suy tuyến tính k-NN) và TabDDPM (mô hình khuếch tán học sâu), đối chứng với kịch bản Baseline không tăng cường. Mục tiêu là làm rõ vai trò của dữ liệu tổng hợp chất lượng cao trong việc cải thiện khả năng phát hiện các lớp tấn công hiếm gặp. Khác với phần lớn các nghiên cứu trước chỉ đánh giá trên dữ liệu đầy đủ, đề tài đặc biệt chú trọng kịch bản nhãn khan hiếm — vốn phản ánh đúng thực tế khi việc thu thập và gán nhãn dữ liệu tấn công hiếm trong môi trường y tế là rất tốn kém và khó khăn.

1.2. Mục tiêu đề tài

Mục tiêu tổng quát của đề tài là đánh giá có hệ thống tác động của các phương pháp tăng cường dữ liệu (SMOTE và TabDDPM) đến hiệu suất phân loại tấn công trên tập dữ liệu CICIoMT2024 trong điều kiện mất cân bằng dữ liệu nghiêm trọng. Các mục tiêu cụ thể bao gồm:

- Xây dựng pipeline xử lý dữ liệu hoàn chỉnh cho CICIoMT2024: tiền xử lý, ánh xạ nhãn, lấy mẫu phân tầng và chuẩn hóa.
- Cài đặt thủ công mô hình khuếch tán TabDDPM bằng PyTorch và áp dụng theo chiến lược tăng cường per-group.
- So sánh khách quan ba kịch bản (Baseline, SMOTE, TabDDPM) trên cùng bộ phân loại và cùng tập kiểm thử cố định.
- Xác định kịch bản mà mô hình khuếch tán phát huy giá trị rõ rệt nhất, từ đó rút ra khuyến nghị thực tiễn cho triển khai IDS trong môi trường y tế.

Khác với nhiều nghiên cứu chỉ dừng lại ở việc báo cáo một con số hiệu năng tổng thể, đề tài hướng tới một sự hiểu biết sâu sắc hơn về điều kiện nào thì phương pháp nào thực sự hiệu quả. Câu hỏi nghiên cứu trung tâm không đơn thuần là "TabDDPM có tốt

hơn SMOTE không?", mà là "TabDDPM tốt hơn trong hoàn cảnh nào, và vì sao?". Cách đặt vấn đề này giúp đề tài tránh được những kết luận hời hợt và mang lại giá trị thực tiễn cao hơn cho người triển khai hệ thống IDS.

1.3. Đối tượng nghiên cứu

- **Tập dữ liệu** — Tập dữ liệu: CICIoMT2024 gồm 51 nhãn con được ánh xạ về 6 nhóm: DDoS, DoS, Reconnaissance, MQTT, Spoofing và Benign.

- **Bộ phân loại** — Dùng bộ phân loại cây RandomForest.

- **Phương pháp tăng cường** — SMOTE (nội suy k-NN) và TabDDPM (mô hình khuếch tán Gaussian).

- **Vấn đề nghiên cứu** — Giải quyết mất cân bằng dữ liệu trong bài toán phân loại đa lớp ở môi trường mạng IoMT.

Nhóm	Vai trò	Đặc điểm về số lượng
Benign	Lưu lượng bình thường	Lớp tham chiếu
DDoS	Từ chối dịch vụ phân tán	Lớp đa số, áp đảo
DoS	Từ chối dịch vụ	Lớp đa số
MQTT	Khai thác giao thức MQTT	Trung bình
Reconnaissance	Trinh sát, dò quét	Lớp thiểu số
Spoofing	Giả mạo ARP	Lớp thiểu số cực đoan

Bảng 1.1. Sáu nhóm phân loại trong phạm vi nghiên cứu

1.4. Phạm vi nghiên cứu

- **Tập dữ liệu** — CICIoMT2024 (gộp 51 nhãn con về 6 nhóm).

- **Mô hình** —Bộ phân loại cây RandomForest, kết hợp hai mô hình sinh dữ liệu (SMOTE và TabDDPM).

- **Triển khai** — Offline (huấn luyện — kiểm thử trên dữ liệu tĩnh). Thực nghiệm chạy trên CPU AMD Ryzen 7 5800H / 16GB RAM và được tái lập trên GPU NVIDIA RTX 3060.

- **Đánh giá** — F1-score cho mỗi nhóm, Macro F1 và Weighted F1.

1.5. Bố cục đồ án

Cấu trúc nội dung của đồ án được quy hoạch đồng bộ và chia thành 5 chương cụ thể như sau:

Chương	Tên chương	Nội dung chính
1	Giới thiệu	Tổng quan, mục tiêu, đối tượng và phạm vi nghiên cứu.
2	Cơ sở lý thuyết	IoMT và an ninh mạng, mất cân bằng dữ liệu, SMOTE, TabDDPM, RF và các chỉ số đánh giá.
3	Thiết kế và triển khai	Tiền xử lý CICIoMT2024, pipeline tăng cường dữ liệu, quy trình huấn luyện bộ phân loại.
4	Kết quả và đánh giá	So sánh các tổ hợp, phân tích per-nhóm, thí nghiệm khan hiếm dữ liệu chứng minh ưu thế TabDDPM.
5	Kết luận và hướng phát triển	Tổng kết đóng góp và đề xuất hướng nghiên cứu tiếp theo.

Bảng 1.2. Cấu trúc nội dung báo cáo

1.6. Phương pháp nghiên cứu

Đề tài kết hợp phương pháp nghiên cứu lý thuyết và phương pháp thực nghiệm. Về lý thuyết, nhóm tiến hành khảo sát tài liệu về an ninh mạng IoMT, các kỹ thuật xử lý mất cân bằng dữ liệu và các mô hình sinh dữ liệu sâu, từ đó xác định khoảng trống nghiên cứu và lựa chọn hướng tiếp cận. Về thực nghiệm, nhóm xây dựng một pipeline đầy đủ trên tập dữ liệu CICIoMT2024, triển khai các mô hình bằng Python (PyTorch, Scikit-learn, imbalanced-learn), tiến hành huấn luyện — kiểm thử có kiểm soát và phân tích định lượng kết quả.

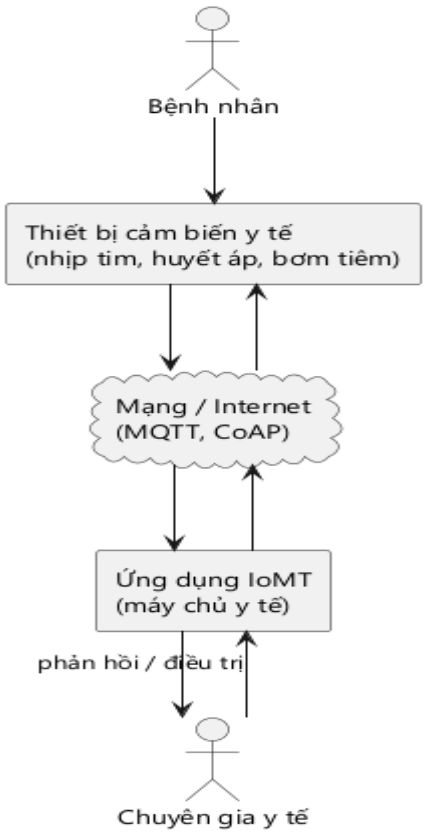
Toàn bộ quá trình tuân thủ nguyên tắc so sánh công bằng và khả năng tái lập: mọi kịch bản dùng chung quy trình tiền xử lý, cùng bộ phân loại với hạt giống ngẫu nhiên cố định, và được đánh giá trên cùng một tập kiểm thử. Cách làm này bảo đảm các kết luận rút ra phản ánh đúng bản chất của phương pháp thay vì các yếu tố ngẫu nhiên.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

2.1. IoMT và an ninh mạng

2.1.1. Định nghĩa và đặc điểm IoMT

Internet of Medical Things (IoMT) là sự kết hợp giữa các thiết bị y tế với Internet of Things (IoT). IoMT được xem là tương lai của các hệ thống y tế hiện nay, nơi mọi thiết bị y tế đều được kết nối và giám sát qua Internet bởi các chuyên gia y tế. Điều này giúp việc chăm sóc sức khỏe trở nên nhanh chóng hơn và giảm chi phí khi công nghệ ngày càng phát triển. Trong mô hình IoMT, các chỉ số sinh tồn (vitals) của bệnh nhân được thu thập qua thiết bị cảm biến và truyền đến các ứng dụng IoMT qua Internet; thông tin sau đó được chuyển tiếp đến nhân viên y tế, từ đó phản hồi/điều trị sẽ được gửi trở lại cho bệnh nhân khi cần thiết.



Hình 2.1. Hệ sinh thái Internet of Medical Things (IoMT)

Về mặt kỹ thuật, các thiết bị IoMT sở hữu ba đặc trưng cốt lõi giúp phân biệt rõ ràng với các thiết bị IoT phổ thông:

- **Tài nguyên phần cứng hạn chế:** Nhằm tối ưu hóa tính cơ động và tiết kiệm điện năng, phần lớn thiết bị IoMT được tích hợp các vi xử lý năng lực thấp cùng dung

lượng bộ nhớ RAM cực kỳ khiêm tốn. Rào cản vật lý này khiến việc cài đặt hay vận hành các kiến trúc bảo mật nặng như tường lửa hay phần mềm diệt virus thế hệ mới trở nên bất khả thi.

- **Giao thức truyền thông đặc thù:** Môi trường IoMT ưu tiên sử dụng các giao thức mạng cấu trúc nhẹ như MQTT (Message Queuing Telemetry Transport) hay CoAP (Constrained Application Protocol). Bản chất của các giao thức này là ưu tiên tối đa cho hiệu suất truyền tải và tiết kiệm băng thông hơn là tích hợp cơ chế mã hóa bảo mật chuyên sâu.

- **Mức độ bảo mật thông tin tối cao:** Các khối dữ liệu luân chuyển trong mạng IoMT thuộc nhóm thông tin định danh y tế nhạy cảm nhất của cá nhân, chịu sự ràng buộc nghiêm ngặt bởi các hành lang pháp lý quốc tế như HIPAA tại Hoa Kỳ hay GDPR tại Liên minh Châu Âu.

Về mặt kiến trúc, một hệ thống IoMT điển hình được tổ chức thành ba tầng có quan hệ chặt chẽ. Tầng cảm nhận (perception layer) gồm các cảm biến và thiết bị đo lường gắn trực tiếp trên hoặc gần bệnh nhân, chịu trách nhiệm thu thập tín hiệu sinh học thô. Tầng mạng (network layer) đảm nhiệm việc truyền tải dữ liệu từ thiết bị tới máy chủ thông qua các giao thức nhẹ như MQTT, CoAP; đây cũng là tầng dễ bị tấn công nhất vì lưu lượng đi qua các kênh không dây và thường thiếu mã hóa đầu-cuối. Tầng ứng dụng (application layer) xử lý, lưu trữ và trực quan hóa dữ liệu cho nhân viên y tế, đồng thời ra quyết định điều trị.

Mỗi tầng phơi bày một bề mặt tấn công riêng: tầng cảm nhận dễ bị giả mạo thiết bị và tiêm dữ liệu sai lệch; tầng mạng là mục tiêu của các đợt flooding và nghe lén; tầng ứng dụng đối mặt với khai thác lỗ hổng phần mềm và đánh cắp dữ liệu. Hệ thống IDS dựa trên học máy trong nghiên cứu này hoạt động chủ yếu ở tầng mạng — nơi có thể quan sát và phân tích đặc trưng luồng lưu lượng để phát hiện hành vi bất thường mà không cần truy cập nội dung dữ liệu y tế nhạy cảm.

2.1.2. Các loại tấn công phổ biến trong CICIoMT2024

Tập dữ liệu an ninh mạng chuẩn hóa CICIoMT2024 phân hóa 18 hình thức xâm nhập mạng vào 5 nhóm hành vi chiến lược, được ánh xạ về 6 nhóm (5 tấn công + Benign) trong phạm vi thực nghiệm này:

- **Nhóm Reconnaissance (trinh sát):** Bao gồm các kỹ thuật quét Port Scan, OS Scan và Recon VulScan. Đây là giai đoạn thám thính ban đầu nhằm dò tìm sơ đồ mạng và phát hiện lỗ hổng dịch vụ trên thiết bị IoMT trước khi kích hoạt tấn công sâu. Đây cũng là nhóm tập trung các lớp thiểu số cực đoan nhất trong tập dữ liệu.

- **Nhóm DoS/DDoS (từ chối dịch vụ):** Bao gồm các biến thể flooding diện rộng như ICMP, TCP, UDP, SYN, Connect Flood và Publish Flood. Mục tiêu là làm tê liệt năng lực tính toán và vắt kiệt băng thông của hạ tầng y tế. Đây là nhóm lớp đa số chiếm tỷ trọng áp đảo.

- **Nhóm Spoofing (giả mạo):** Đại diện tiêu biểu là kỹ thuật ARP Spoofing, trong đó tác nhân đe dọa giả mạo địa chỉ phần cứng của các thực thể hợp lệ nhằm đánh chặn, điều hướng hoặc làm sai lệch dòng thông tin y khoa.

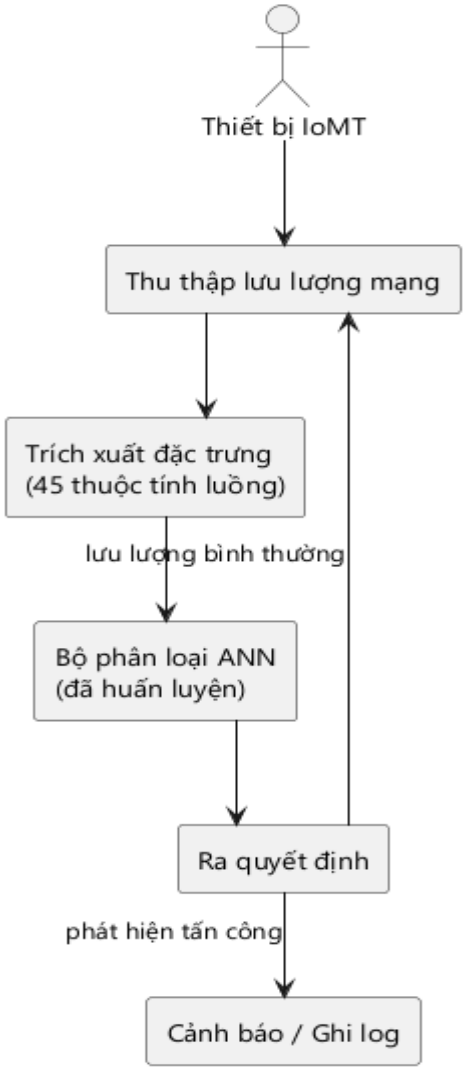
- **Nhóm MQTT-based (khai thác giao thức):** Bao gồm DDoS/DoS Connect Flood, DDoS/DoS Publish Flood và Malformed Data. Đây là nhóm hiểm họa đặc thù của không gian IoMT, tấn công trực diện vào điểm yếu của giao thức MQTT vốn đang được ứng dụng rộng rãi.

Về mặt cơ chế, mỗi nhóm tấn công khai thác một điểm yếu khác nhau của hạ tầng IoMT. Tấn công DDoS/DoS dạng flooding gửi một lượng khổng lồ gói tin (TCP SYN, UDP, ICMP) trong thời gian ngắn nhằm vắt kiệt tài nguyên xử lý và băng thông, khiến thiết bị giám sát bệnh nhân không kịp phản hồi các yêu cầu hợp lệ. Tấn công ARP Spoofing lợi dụng việc giao thức ARP không có cơ chế xác thực: kẻ tấn công phát đi các bản tin ARP giả mạo để liên kết địa chỉ MAC của mình với địa chỉ IP của thiết bị hợp lệ, từ đó chen vào giữa luồng dữ liệu (man-in-the-middle) để nghe lén hoặc sửa đổi thông tin y khoa.

Nhóm tấn công khai thác giao thức MQTT đặc biệt nguy hiểm vì MQTT là xương sống truyền thông của nhiều thiết bị IoMT. Connect Flood liên tục mở các kết nối CONNECT giả để làm cạn kiệt bộ nhớ của MQTT broker; Publish Flood spam các bản tin PUBLISH làm nghẽn hàng đợi thông điệp; còn Malformed Data gửi các gói tin dị dạng nhằm khai thác lỗi xử lý của broker. Trong khi đó, nhóm Reconnaissance không trực tiếp gây thiệt hại mà âm thầm thu thập thông tin (quét cổng, quét hệ điều hành, dò lỗ hổng) để chuẩn bị cho các đợt tấn công sâu hơn — đây cũng là nhóm khó phát hiện nhất do lưu lượng thấp và lén lút.

2.1.3. Hệ thống phát hiện xâm nhập (IDS)

Hệ thống phát hiện xâm nhập (IDS) đóng vai trò là một chốt chặn an ninh then chốt, thực hiện nhiệm vụ giám sát liên tục toàn bộ lưu lượng mạng nhằm nhận diện và phát cảnh báo trước các hành vi bất thường hoặc có dấu hiệu độc hại. Kiến trúc IDS hiện nay được phân định theo hai hướng tiếp cận nền tảng: IDS dựa trên dấu hiệu (Signature-based) và IDS dựa trên học máy (ML-based).



Hình 2.2. Mô hình chung của hệ thống IDS dựa trên học máy

IDS dựa trên ML học đặc trưng thống kê từ lưu lượng mạng thực tế. Khác với Signature-based IDS bị giới hạn bởi cơ sở dữ liệu tấn công đã biết, ML-IDS có khả năng phát hiện tấn công biến thể và zero-day [5]. Tuy nhiên, hiệu quả của ML-IDS phụ thuộc trực tiếp vào chất lượng và sự cân bằng của dữ liệu huấn luyện [6, 7].

IDS dựa trên dấu hiệu hoạt động bằng cách so khớp lưu lượng với một cơ sở dữ liệu các mẫu tấn công đã biết (signature). Ưu điểm của nó là độ chính xác cao và tỷ lệ

báo động giả thấp đối với các tấn công đã được định nghĩa; nhược điểm chí mạng là hoàn toàn "mù" trước các tấn công mới hoặc biến thể chưa có trong cơ sở dữ liệu. Trong bối cảnh các kỹ thuật tấn công IoMT liên tục tiến hóa, hạn chế này khiến IDS dựa trên dấu hiệu trở nên không đủ. Ngược lại, IDS dựa trên học máy học các đặc trưng thống kê khái quát của hành vi bình thường và bất thường, nhờ đó có khả năng phát hiện cả những tấn công chưa từng gặp — đây chính là lý do đề tài lựa chọn hướng tiếp cận học máy.

2.2. Mất cân bằng dữ liệu (Class Imbalance)

2.2.1. Định nghĩa và nguyên nhân

Mất cân bằng dữ liệu là sự phân bố bất đối xứng cực đoan giữa các lớp trong tập huấn luyện [6]. Trong CICIoMT2024 [1], sau khi lấy mẫu phân tầng, tập huấn luyện vẫn lệch nghiêm trọng: nhóm DDoS chiếm hàng trăm nghìn mẫu trong khi nhóm Spoofing (ARP) chỉ có khoảng 16.000 mẫu — và sau khi mô phỏng kịch bản khan hiếm, chênh lệch còn lớn hơn nữa. Nguyên nhân gốc rễ đến từ bản chất lưu lượng mạng: hành vi bình thường và các đợt flooding diện rộng tạo ra khối lượng gói tin khổng lồ, trong khi các hành vi trinh sát hay giả mạo lại tinh vi, lén lút và hiếm gặp.

Trong lĩnh vực an ninh mạng nói chung và IoMT nói riêng, mất cân bằng dữ liệu không phải là một khiếm khuyết có thể dễ dàng loại bỏ, mà là một đặc tính nội tại của bài toán. Bản chất của an ninh mạng là phần lớn lưu lượng đều bình thường, còn các cuộc tấn công — đặc biệt là những loại tinh vi và nguy hiểm nhất — vốn dĩ hiếm gặp. Càng là tấn công nguy hiểm thì càng ít xuất hiện trong dữ liệu thu thập được, tạo nên một nghịch lý: chính những lớp cần được phát hiện nhất lại là những lớp khó học nhất. Đây là lý do vì sao xử lý mất cân bằng dữ liệu trở thành một trong những chủ đề trung tâm và bền bỉ nhất của nghiên cứu về IDS dựa trên học máy.

Mức độ mất cân bằng thường được định lượng qua tỷ số mất cân bằng (Imbalance Ratio — IR), bằng số mẫu lớp đa số chia cho số mẫu lớp thiểu số. Một tập dữ liệu được coi là mất cân bằng nhẹ khi IR dưới 10, mất cân bằng trung bình khi IR từ 10 đến 100, và mất cân bằng nghiêm trọng khi IR vượt 100. Với CICIoMT2024, ngay cả sau khi lấy mẫu phân tầng, tỷ số giữa nhóm DDoS và nhóm Spoofing đã vào khoảng 43; còn trong kịch bản khan hiếm dữ liệu mô phỏng ở Chương 4, tỷ số này còn cao hơn nhiều khi

nhóm Spoofing chỉ còn 200 mẫu. Đây là mức độ mất cân bằng thuộc loại khắc nghiệt, đủ để vô hiệu hóa các bộ phân loại thông thường nếu không có biện pháp can thiệp.

2.2.2. Tác hại đối với mô hình học máy

Khi một lớp chiếm tỉ lệ áp đảo trong dữ liệu huấn luyện, mô hình có xu hướng học theo lớp đa số để tối ưu hóa độ chính xác tổng thể (accuracy), vì dự đoán "luôn là lớp đa số" cũng đã cho accuracy rất cao. Hệ quả là mô hình gần như không học được đặc trưng của lớp thiểu số.

Các lớp tấn công hiếm thường có recall rất thấp, nghĩa là mô hình không nhận diện được khi tấn công đó xảy ra trong thực tế, dẫn đến bị phân loại nhầm thành benign hoặc một loại tấn công khác. Trong an ninh mạng, "bỏ sót tấn công" nguy hiểm hơn nhiều so với "báo động giả".

Hiện tượng này thường được gọi là "nghịch lý độ chính xác" (accuracy paradox): một mô hình có thể đạt độ chính xác tổng thể rất cao nhưng lại hoàn toàn vô dụng trong thực tế. Ví dụ, nếu lớp tấn công chỉ chiếm 1% dữ liệu, một mô hình luôn dự đoán "bình thường" đã đạt độ chính xác 99% — nhưng bỏ sót 100% các cuộc tấn công. Trong an ninh mạng y tế, đây là kịch bản tồi tệ nhất có thể xảy ra. Chính vì vậy, các chỉ số đánh giá cần được lựa chọn cẩn trọng để phản ánh đúng năng lực phát hiện trên lớp thiểu số, thay vì bị đánh lừa bởi độ chính xác tổng thể — vấn đề sẽ được phân tích kỹ ở mục 2.5.

2.2.3. Các hướng giải quyết

Các giải pháp kỹ thuật nhằm chế ngự bài toán mất cân bằng phân phối trong học máy thường được phân định thành ba trường phái chính:

- **Cấp độ dữ liệu (Data-level):** Tập trung vào việc tái cấu trúc phân phối mẫu thông qua undersampling (loại bớt mẫu lớp đa số) hoặc oversampling (gia tăng mẫu lớp thiểu số). Phương pháp oversampling dựa trên sinh dữ liệu tổng hợp được đánh giá cao hơn nhờ bảo toàn trọn vẹn thông tin gốc.
- **Cấp độ thuật toán (Algorithm-level):** Can thiệp trực tiếp vào hàm mục tiêu bằng các kỹ thuật như điều chỉnh trọng số hàm mất mát theo lớp (class-weighted loss) hay học nhạy cảm chi phí (cost-sensitive learning).
- **Hỗn hợp (Hybrid):** Đồng bộ hóa ưu thế của cả hai giải pháp trên nhằm đạt hiệu năng tối ưu nhất.

Nghiên cứu này tập trung vào hướng tiếp cận cấp độ dữ liệu, so sánh trực tiếp SMOTE và TabDDPM.

Lý do của sự lựa chọn này là tính linh hoạt và khả năng tổng quát của hướng tiếp cận cấp độ dữ liệu: các phương pháp tăng cường hoạt động độc lập với bộ phân loại, nên một khi tạo ra được tập dữ liệu cân bằng và chất lượng, tập đó có thể dùng cho bất kỳ mô hình phân loại nào. Điều này khác biệt với hướng cấp độ thuật toán, vốn gắn chặt với một kiến trúc cụ thể và khó chuyển giao. Trong bối cảnh nghiên cứu so sánh, tính tách biệt giữa giai đoạn tăng cường và giai đoạn phân loại còn giúp cô lập rõ ràng tác động của từng phương pháp tăng cường, phục vụ cho việc đánh giá khách quan và công bằng.

2.3. Các phương pháp xử lý mất cân bằng được so sánh

2.3.1. Baseline (Không tăng cường)

Trong kịch bản Baseline, bộ phân loại được huấn luyện trực tiếp trên tập sau tiền xử lý, không có bất kỳ augmentation nào. Đây là hệ quy chiếu để định lượng chính xác phần cải tiến mà SMOTE hay TabDDPM mang lại.

2.3.2. SMOTE — Synthetic Minority Over-sampling Technique

SMOTE [8] tạo mẫu tổng hợp bằng nội suy tuyến tính trong không gian đặc trưng. Với mỗi mẫu thuộc nhóm thiểu số, thuật toán chọn ngẫu nhiên một trong các láng giềng gần nhất cùng nhóm và tạo mẫu mới theo công thức:

$$x_{new} = x_i + \lambda(x_{nn} - x_i), \quad \lambda \sim \mathcal{U}(0,1)$$

Quy trình của SMOTE có thể tóm tắt qua bốn bước: (1) với mỗi mẫu thiểu số, tìm k láng giềng gần nhất cùng lớp; (2) chọn ngẫu nhiên một láng giềng trong số đó; (3) sinh điểm mới nằm trên đoạn thẳng nối hai điểm theo hệ số λ ngẫu nhiên; (4) lặp lại cho đến khi đạt số mẫu mục tiêu. Bằng cách này, SMOTE mở rộng vùng quyết định của lớp thiểu số mà không đơn thuần nhân bản mẫu cũ.

trong đó λ là hệ số nội suy ngẫu nhiên trong khoảng (0, 1). Ngưỡng mục tiêu số mẫu của mỗi nhóm được tính thống nhất với TabDDPM theo công thức phân đoạn:

$$n_{target} = \min(3\,n_{real},\ 200000) \quad \text{khi} \ \ n_{real} < 200000$$
$$n_{target} = 200000 + 0.2\,(n_{real} - 200000) \quad \text{khi} \ \ 200000 \le n_{real} \le 500000$$
$$n_{target} = 360000 + 0.1\,(n_{real} - 500000) \quad \text{khi} \ \ n_{real} > 500000$$

SMOTE được áp dụng theo chiến lược per-group qua từ điển sampling_strategy, chỉ bao gồm các nhóm chưa đạt mục tiêu. Tham số k_neighbors được điều chỉnh an toàn để tránh lỗi với nhóm siêu nhỏ, triển khai qua thư viện imbalanced-learn.

Tham số	Giá trị
k_neighbors	5 (tự điều chỉnh theo nhóm nhỏ nhất)
SMALL_THR_SMOTE	200.000
MEDIUM_THR_SMOTE	500.000
random_state	42

Bảng 2.1. Siêu tham số SMOTE

Ưu điểm của SMOTE là nhanh, không cần huấn luyện mô hình tham số, không có rủi ro phân kỳ. Tuy nhiên, SMOTE chỉ nội suy tuyến tính trong bao lồi (convex hull) nên không nắm bắt được phân phối phi tuyến phức tạp. Trong không gian 45 chiều, khoảng cách Euclidean của k-NN chịu "lời nguyền chiều không gian" [16] — chất lượng láng giềng suy giảm; các mẫu biên giữa Reconnaissance và Benign dễ tạo ra mẫu nhiễu.

"Lời nguyền chiều không gian" (curse of dimensionality) là một hiện tượng nền tảng trong học máy: khi số chiều tăng, các điểm dữ liệu trở nên thưa thớt và khoảng cách giữa chúng có xu hướng đồng đều hóa, khiến khái niệm "láng giềng gần nhất" mất dần ý nghĩa. Với 45 chiều như trong CICIoMT2024, hai mẫu gần nhau theo khoảng cách Euclidean chưa chắc đã thực sự tương đồng về mặt hành vi mạng. Hệ quả là các mẫu nội suy của SMOTE — vốn được tạo trên đoạn thẳng nối hai điểm "láng giềng" — có thể rơi vào vùng không thuộc về phân phối thực của lớp, trở thành mẫu nhiễu. Đây là hạn chế cố hữu của các phương pháp dựa trên khoảng cách mà mô hình khuếch tán, nhờ học toàn bộ phân phối thay vì chỉ quan hệ cục bộ giữa các điểm, có thể khắc phục được.

2.3.3. TabDDPM — Denoising Diffusion Probabilistic Model

TabDDPM [9] là mô hình khuếch tán xác suất thuần Gaussian (PureGaussianDiffusion) được triển khai thủ công bằng PyTorch, gồm hai pha đối xứng. Pha xuôi (forward) tiêm nhiễu Gaussian tăng dần qua T bước theo lịch trình tuyến tính:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right)$$

Nhờ tính chất Markov, có thể lấy mẫu trực tiếp tại bước t bất kỳ theo dạng đóng:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$$

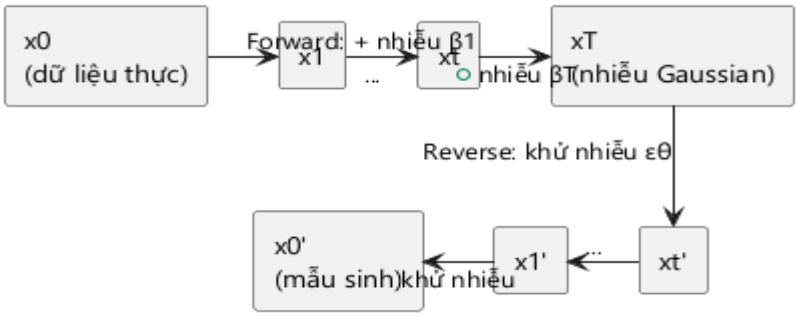
Về mặt trực giác, có thể hình dung pha xuôi như quá trình một bức ảnh rõ nét bị phủ nhiễu dần cho đến khi trở thành nhiễu trắng hoàn toàn; còn pha ngược là quá trình học cách "tẩy nhiễu" để tái tạo lại bức ảnh. Khi áp dụng cho dữ liệu dạng bảng, mỗi "bức ảnh" là một vector đặc trưng của một mẫu lưu lượng mạng. Sau khi học được cách khử nhiễu, mô hình có thể sinh ra mẫu mới hoàn toàn bằng cách bắt đầu từ nhiễu ngẫu nhiên và khử dần — qua đó tạo ra dữ liệu tổng hợp đa dạng nhưng vẫn tuân theo phân phối của lớp gốc.

Pha ngược (reverse) sử dụng mạng nơ-ron học cách khử nhiễu từng bước để khôi phục dữ liệu:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\right)$$

Hàm mục tiêu huấn luyện là sai số dự đoán nhiễu (noise prediction):

$$\mathcal{L}_{\text{sml}} = E_{x_0, t, \varepsilon}[|\varepsilon - \varepsilon_\theta(x_t, t)|^2]$$



Hình 2.3. Pha xuôi (thêm nhiễu) và pha ngược (khử nhiễu) của mô hình khuếch tán

Kiến trúc mạng khử nhiễu TimeConditionedDenoiser bao gồm: bộ mã hóa thời gian SinusoidalTimeEmb để biến đổi bước thời gian t thành embedding, cộng với phép chiếu đặc trưng đầu vào; các khối DenoiserBlock (Linear → BatchNorm → SiLU → Dropout) sử dụng hàm kích hoạt SiLU [14] thay cho ReLU vì tính trơn và đạo hàm liên tục giúp gradient flow tốt hơn; backbone gồm nhiều khối với chiều giảm dần.

Để tương thích với giả định phân phối Gaussian, dữ liệu thô của mỗi nhóm được fit riêng bằng QuantileTransformer (output_distribution = normal, n_quantiles = 1000).

Sau khi sinh mẫu, kết quả được cắt biên về khoảng [-5, 5] rồi inverse_transform về không gian gốc. Chất lượng mẫu được kiểm soát hai tầng: thứ nhất là độ lệch trung bình cực đại MMD (Maximum Mean Discrepancy) đo khoảng cách phân phối giữa mẫu thực và mẫu tổng hợp bằng RBF kernel:

$$\mathrm{MMD}^2 = \frac{1}{n^2}\sum_{i,j} k(x_i, x_j) + \frac{1}{m^2}\sum_{i,j} k(y_i, y_j) - \frac{2}{nm}\sum_{i,j} k(x_i, y_j)$$

Nếu MMD vượt ngưỡng, toàn bộ mẫu sinh của nhóm đó bị loại bỏ. Tầng thứ hai là bộ lọc IsolationForest: huấn luyện trên dữ liệu thực rồi chấm điểm độ "giống thật" của mẫu sinh, giữ lại các mẫu chất lượng cao nhất và loại bỏ các mẫu dị thường.

Tham số	Giá trị
T_STEPS (số bước khuếch tán)	500 – 1000
BETA (lịch nhiễu)	1e-4 → 0,02 (tuyến tính)
DIM_T (chiều embedding thời gian)	128 – 256
LR_DDPM	1e-3 (AdamW + CosineAnnealing)
EPOCHS	300 – 600 (thích nghi theo quy mô nhóm)
MMD_THRESHOLD	0,5
IF_KEEP_FRAC	0,90

Bảng 2.2. Siêu tham số TabDDPM

Ưu điểm của TabDDPM là nắm bắt phân phối phi tuyến phức tạp, chất lượng mẫu được kiểm soát chặt chẽ. Nhược điểm là chi phí tính toán lớn: mỗi nhóm cần hàng trăm epoch huấn luyện và quá trình lấy mẫu tuần tự qua T bước cho mỗi batch.

Một câu hỏi tự nhiên là tại sao mô hình khuếch tán — vốn được phát triển ban đầu cho ảnh — lại phù hợp với dữ liệu mạng dạng bảng. Lý do nằm ở tính tổng quát của khung lý thuyết: bản chất của khuếch tán là học cách biến đổi giữa phân phối dữ liệu phức tạp và phân phối Gaussian đơn giản, không phụ thuộc vào việc dữ liệu là ảnh, âm thanh hay vector đặc trưng. Với dữ liệu bảng, mỗi đặc trưng đóng vai trò như một "chiều" trong không gian; sau khi được biến đổi về phân phối chuẩn bằng QuantileTransformer, dữ liệu hoàn toàn tương thích với giả định Gaussian của mô hình. Chính sự linh hoạt này khiến TabDDPM trở thành một lựa chọn mạnh mẽ cho bài toán sinh dữ liệu tổng hợp trong an ninh mạng.

Quá trình lấy mẫu (sampling) của mô hình khuếch tán được thực hiện tuần tự theo chiều ngược thời gian, bắt đầu từ nhiễu Gaussian thuần túy và khử dần qua T bước. Thuật toán được tóm lược như sau:

```
Thuat toan: Lay mau tu mo hinh khuech tan
  Dau vao : so mau n, so chieu d, mang khu nhieu eps_theta
  1. x_T  <- N(0, I)            # khoi tao tu nhieu Gaussian
  2. for t = T, T-1, ..., 1:
  3.     eps  <- eps_theta(x_t, t)         # du doan nhieu
  4.     x0_hat <- (x_t - sqrt(1-acp_t)*eps) / sqrt(acp_t)
  5.     x0_hat <- clip(x0_hat, -10, 10)
  6.     mu  <- pc1_t * x0_hat + pc2_t * x_t   # trung binh hau nghiem
  7.     x_{t-1} <- mu + sigma_t * z,  z ~ N(0,I) neu t>1
  8. return x_0                  # mau tong hop
```

2.3.4. Tổng quan và so sánh các mô hình tạo sinh

Có hai họ mô hình tạo sinh sâu phổ biến cho dữ liệu dạng bảng: mô hình khuếch tán (Diffusion) và mô hình nội suy tuyến tính (SMOTE). Mỗi họ có ưu nhược điểm riêng khi áp dụng cho bài toán tăng cường dữ liệu mất cân bằng.

Dựa trên phân tích trên, đề tài lựa chọn TabDDPM làm đại diện cho hướng tiếp cận học sâu tạo sinh, đối chứng với SMOTE (đại diện cho nội suy cổ điển) và Baseline (không tăng cường).

Cần lưu ý rằng việc áp dụng các mô hình tạo sinh cho dữ liệu dạng bảng phức tạp hơn so với dữ liệu ảnh, do dữ liệu bảng thường chứa hỗn hợp các kiểu đặc trưng (liên tục, rời rạc, nhị phân) với phân phối không chuẩn và các ràng buộc logic giữa các cột. Trong nghiên cứu này, toàn bộ 45 đặc trưng đều ở dạng số, và bước biến đổi QuantileTransformer đóng vai trò cầu nối quan trọng: nó chuẩn hóa phân phối của từng đặc trưng về dạng gần Gaussian, giúp mô hình khuếch tán hoạt động hiệu quả trên không gian đặc trưng vốn có phân phối rất lệch của lưu lượng mạng. Đây là một mắt xích kỹ thuật mà nếu thiếu, chất lượng mẫu sinh sẽ suy giảm đáng kể.

2.4. Các chỉ số (Metrics) đánh giá

2.4.1. Các đại lượng cơ bản từ ma trận nhầm lẫn

Với từng lớp, ma trận nhầm lẫn (confusion matrix) được phân rã thành bốn thành phần: TP (số mẫu thuộc lớp được phân loại đúng), FP (số mẫu không thuộc lớp bị nhầm sang lớp đó), FN (số mẫu thuộc lớp bị phân loại sang lớp khác — bỏ lọt), và TN (số mẫu không thuộc lớp và được phân loại đúng là không phải).

Bốn thành phần này là nền tảng để xây dựng mọi chỉ số đánh giá. Trong bối cảnh an ninh mạng, mỗi thành phần mang một ý nghĩa nghiệp vụ cụ thể: TP là các cuộc tấn công được phát hiện đúng; FP là các báo động giả gây phiền nhiễu cho đội ngũ vận hành; FN là các cuộc tấn công bị bỏ lọt — thành phần nguy hiểm nhất vì đồng nghĩa với việc hệ thống thất bại trong nhiệm vụ chính; còn TN là lưu lượng bình thường được nhận diện đúng. Một hệ thống IDS lý tưởng phải tối đa hóa TP, tối thiểu hóa FN, đồng thời giữ FP ở mức chấp nhận được để không gây "mệt mỏi cảnh báo" (alert fatigue) cho người vận hành.

2.4.2. Precision, Recall và F1-score

Precision và Recall được định nghĩa như sau:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score là trung bình điều hòa của Precision và Recall, cân bằng giữa hai chỉ số này:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Trong bối cảnh an ninh mạng IoMT, Recall thấp (tức FN lớn) đồng nghĩa với việc bỏ lọt tấn công thực tế — hậu quả trực tiếp đến an toàn bệnh nhân. Vì vậy F1 là chỉ số cân bằng phù hợp cho bài toán này.

2.4.3. Macro F1-score — Chỉ số cốt lõi

Macro F1 lấy trung bình cộng F1 của mọi lớp, gán trọng số ngang bằng cho mọi nhóm — lớp thiểu số quan trọng như lớp đa số:

$$\text{Macro F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c$$

Ngược lại, Weighted F1 gán trọng số theo số lượng mẫu, nên bị che mờ bởi lớp đa số:

$$\text{Weighted F1} = \frac{1}{N}\sum_{c=1}^{C} n_c \cdot F1_c$$

Macro F1 là chỉ số then chốt cho IDS trong môi trường y tế, vì một lớp tấn công nguy hiểm bị bỏ sót hoàn toàn vẫn không làm giảm Weighted F1 đáng kể [7]. Do đó, đề tài sử dụng Macro F1 làm chỉ số đánh giá chính, kết hợp Weighted F1 để tham chiếu.

2.4.4. Ví dụ minh họa sự khác biệt Macro và Weighted F1

Để thấy rõ vì sao Macro F1 phù hợp hơn cho bài toán mất cân bằng, xét một ví dụ giả định gồm hai lớp: lớp đa số "Bình thường" (9.800 mẫu) và lớp thiểu số "Tấn công" (200 mẫu). Giả sử mô hình A phân loại đúng toàn bộ lớp đa số nhưng chỉ phát hiện được 20 trên 200 mẫu tấn công.

Lớp	Số mẫu	Precision	Recall	F1
Bình thường	9.800	0,982	1,000	0,991
Tấn công	200	1,000	0,100	0,182
Macro (trung bình)	—	0,991	0,550	0,586
Weighted (theo mẫu)	—	0,982	0,982	0,975

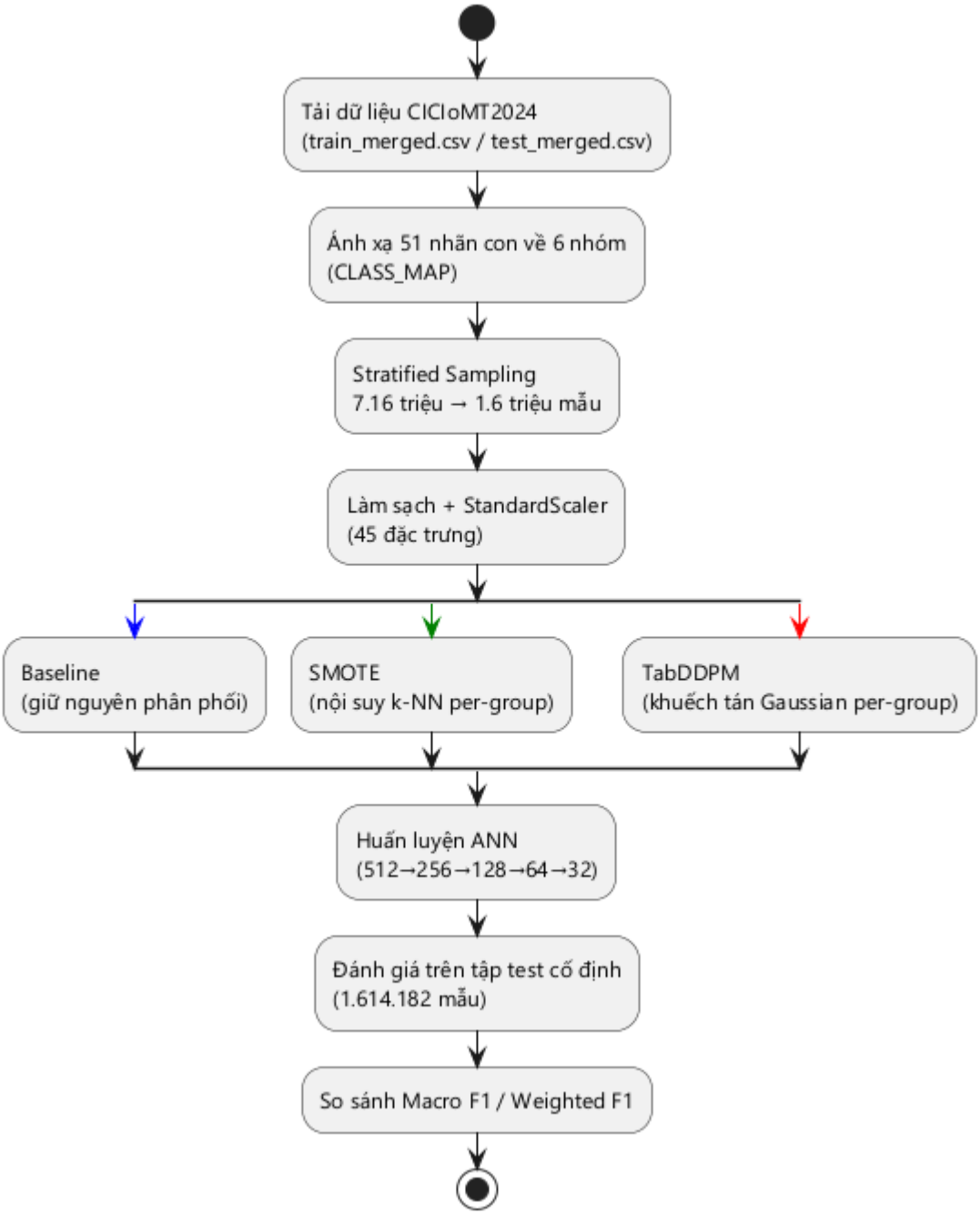
Bảng 2.3. Ví dụ minh họa Macro F1 và Weighted F1

Trong ví dụ trên, Weighted F1 đạt tới 0,975 — một con số rất "đẹp" che giấu hoàn toàn thực tế rằng mô hình bỏ sót 90% các cuộc tấn công. Ngược lại, Macro F1 chỉ đạt 0,586, phản ánh trung thực sự yếu kém trên lớp thiểu số. Đây chính là lý do đề tài lấy Macro F1 làm thước đo chủ đạo: trong an ninh mạng y tế, việc bỏ sót dù chỉ một loại tấn công cũng có thể gây hậu quả nghiêm trọng, nên không thể để chỉ số tổng hợp bị chi phối bởi lớp đa số.

CHƯƠNG 3: THIẾT KẾ VÀ TRIỂN KHAI

3.1. Tổng quan kiến trúc hệ thống

Hệ thống phát hiện xâm nhập (IDS) trong nghiên cứu này được xây dựng theo một quy trình xử lý dữ liệu khép kín và đồng bộ (pipeline), luân chuyển từ các khối dữ liệu thô ban đầu cho đến các mô hình đánh giá hiệu năng cuối cùng. Tiến trình này bao gồm bốn giai đoạn chiến lược: (1) tiền xử lý dữ liệu, (2) tăng cường dữ liệu theo từng kịch bản, (3) huấn luyện bộ phân loại, và (4) đánh giá trên tập kiểm thử cố định.



Hình 3.1. Sơ đồ pipeline thực nghiệm tổng thể

Điểm mấu chốt của thiết kế là tính công bằng trong so sánh: cả ba kịch bản (Baseline, SMOTE, TabDDPM) đều dùng chung một quy trình tiền xử lý, cùng một kiến trúc bộ phân loại với cùng siêu tham số, và được đánh giá trên cùng một tập kiểm thử thực tế. Nhờ đó, mọi khác biệt về hiệu năng đều phản ánh đúng giá trị của phương pháp tăng cường dữ liệu.

Pipeline được tổ chức theo nguyên tắc module hóa: mỗi giai đoạn là một khối chức năng độc lập, có đầu vào và đầu ra rõ ràng, cho phép thay thế hoặc mở rộng mà không ảnh hưởng đến các giai đoạn khác. Tính module hóa này không chỉ thuận lợi cho việc

mở rộng nghiên cứu trong tương lai mà còn bảo đảm khả năng tái lập và kiểm chứng độc lập từng thành phần.

3.2. Quy trình tiền xử lý dữ liệu

3.2.1. Tải dữ liệu và tham số cơ bản

- Tập huấn luyện: ./data/CICIoMT2024/train_merged.csv (≈ 7,16 triệu mẫu).
- Tập kiểm thử: ./data/CICIoMT2024/test_merged.csv (1.614.182 mẫu, cố định cho cả ba kịch bản).
- MIN_SAMPLES = 100: loại bỏ các nhóm quá thưa thớt sau khi lấy mẫu.

3.2.2. Ánh xạ nhãn (CLASS_MAP)

Từ điển CLASS_MAP ánh xạ 51 nhãn con trong CICIoMT2024 về 6 nhóm nghiên cứu. Việc gộp nhãn giúp giảm độ phân mảnh và tập trung vào các nhóm hành vi tấn công có ý nghĩa chiến lược:

```
CLASS_MAP = {
    'Benign'              : 'Benign',
    'ARP_Spoofing'        : 'Spoofing',
    'Recon-OS_Scan'       : 'Reconnaissance',
    'Recon-Port_Scan'     : 'Reconnaissance',
    'Recon-VulScan'       : 'Reconnaissance',
    'MQTT-DDoS-Connect_Flood' : 'MQTT',
    'MQTT-Malformed_Data' : 'MQTT',
    'TCP_IP-DDoS-TCP1'    : 'DDoS',   # ... các biến thể DDoS
    'TCP_IP-DoS-SYN1'     : 'DoS',    # ... các biến thể DoS
}
```

Nhóm	Các nhãn con được gộp
Benign	Benign
Spoofing	ARP_Spoofing
Reconnaissance	Recon-OS_Scan, Recon-Ping_Sweep, Recon-Port_Scan, Recon-VulScan
MQTT	MQTT-DDoS/DoS-Connect_Flood, MQTT-DDoS/DoS-Publish_Flood, MQTT-Malformed_Data
DDoS	TCP_IP-DDoS-TCP / UDP / ICMP / SYN (các biến thể)
DoS	TCP_IP-DoS-TCP / UDP / ICMP / SYN (các biến thể)

Bảng 3.1. Bảng ánh xạ CLASS_MAP từ CICIoMT2024

3.2.3. Lấy mẫu phân tầng (Stratified Sampling)

Do tập huấn luyện gốc rất lớn (≈ 7,16 triệu dòng), nhóm áp dụng chiến lược lấy mẫu phân tầng để giảm khối lượng tính toán mà vẫn giữ nguyên tính đại diện thống kê. Chiến lược giữ lại toàn bộ nhóm nhỏ, đồng thời cắt giảm có kiểm soát các nhóm đa số:

Quy mô nhóm gốc	Chiến lược	Mô tả
< 200.000	Giữ toàn bộ	Bảo toàn hoàn toàn nhóm nhỏ và thiểu số
200.000 – 500.000	Cắt có kiểm soát	Giữ 200k + 20% phần vượt
> 500.000	Cắt tích lũy	Giữ 360k + 10% phần vượt

Bảng 3.2. Chiến lược lấy mẫu phân tầng trên CICIoMT2024

Sau khi lấy mẫu, tập huấn luyện giảm xuống còn khoảng 1,6 triệu dòng (tiết kiệm gần 78% khối lượng) nhưng vẫn bảo toàn sự phân tầng giữa các nhóm.

3.2.4. Làm sạch và chuẩn hóa

- Chỉ giữ các cột số (float64, int64, float32, int32).
- Loại bỏ cột Drate (100% giá trị NaN), lấp các giá trị NaN còn lại bằng 0. Kết quả thu được 45 đặc trưng hợp lệ.
- Áp dụng StandardScaler (fit trên tập train, transform cả train và test) để chuẩn hóa về zero-mean unit-variance, ngăn rò rỉ dữ liệu (data leakage) theo thực hành chuẩn [15, 16].
- Riêng cho TabDDPM: lưu thêm bản dữ liệu chưa chuẩn hóa để dùng cho QuantileTransformer theo từng nhóm.

Việc chuẩn hóa đóng vai trò quan trọng đối với cả bộ phân loại lẫn mô hình sinh dữ liệu. Các đặc trưng luồng mạng có thang đo rất khác nhau — ví dụ, kích thước gói tính bằng byte có thể lên đến hàng nghìn, trong khi các cờ giao thức chỉ nhận giá trị 0 hoặc 1. Nếu không chuẩn hóa, các đặc trưng có thang đo lớn sẽ áp đảo quá trình học, làm sai lệch trọng số của mạng nơ-ron và khoảng cách trong không gian đặc trưng. StandardScaler đưa mọi đặc trưng về cùng thang đo (trung bình 0, độ lệch chuẩn 1), bảo đảm sự công bằng giữa các đặc trưng. Điều quan trọng là bộ chuẩn hóa chỉ được "học" (fit) trên tập huấn luyện rồi áp dụng cho tập kiểm thử, tránh rò rỉ thông tin từ tập test vào quá trình huấn luyện — một lỗi nghiêm trọng nhưng phổ biến trong thực hành học máy.

Bộ 45 đặc trưng số được trích xuất từ luồng mạng, có thể phân thành các nhóm theo ý nghĩa thống kê như mô tả trong Bảng 3.3. Đây đều là các đặc trưng cấp luồng (flow-level), không chứa nội dung gói tin (payload), nên bảo toàn quyền riêng tư của dữ liệu y tế.

Nhóm đặc trưng	Ví dụ	Ý nghĩa
Thống kê kích thước gói	Header_Length, Tot size, Min, Max, AVG, Std	Phân bố độ lớn gói tin trong luồng
Tốc độ và thời gian	Rate, Srate, Drate, IAT, Duration	Tần suất và khoảng cách thời gian giữa các gói
Cờ giao thức TCP	syn_flag, ack_flag, fin_flag, rst_flag	Trạng thái bắt tay và điều khiển kết nối
Loại giao thức	TCP, UDP, ICMP, HTTP, MQTT, ARP	Phân loại giao thức tầng mạng/ứng dụng
Thống kê dòng	Number, Magnitude, Radius, Covariance, Variance	Đặc trưng phân bố tổng hợp của luồng

Bảng 3.3. Phân nhóm các đặc trưng đầu vào

Bảng 3.4 liệt kê đầy đủ 45 đặc trưng đầu vào cùng ý nghĩa của chúng. Đây là tập đặc trưng chuẩn của CICIoMT2024, được trích xuất hoàn toàn ở cấp luồng mạng.

STT	Tên đặc trưng	Ý nghĩa
1	Header_Length	Độ dài phần tiêu đề gói tin
2	Protocol Type	Loại giao thức tầng vận chuyển
3	Duration	Thời lượng tồn tại của luồng
4	Rate	Tốc độ gói tin tổng thể
5	Srate	Tốc độ gói tin chiều gửi (source)
6	Drate	Tốc độ gói tin chiều nhận (destination)
7	fin_flag_number	Số cờ FIN (kết thúc kết nối)
8	syn_flag_number	Số cờ SYN (thiết lập kết nối)
9	rst_flag_number	Số cờ RST (đặt lại kết nối)
10	psh_flag_number	Số cờ PSH (đẩy dữ liệu)
11	ack_flag_number	Số cờ ACK (xác nhận)
12	ece_flag_number	Số cờ ECE (thông báo tắc nghẽn)
13	cwr_flag_number	Số cờ CWR (giảm cửa sổ tắc nghẽn)
14	ack_count	Tổng số gói ACK
15	syn_count	Tổng số gói SYN
16	fin_count	Tổng số gói FIN
17	rst_count	Tổng số gói RST

18	HTTP	Cờ giao thức HTTP
19	HTTPS	Cờ giao thức HTTPS
20	DNS	Cờ giao thức DNS
21	Telnet	Cờ giao thức Telnet
22	SMTP	Cờ giao thức SMTP
23	SSH	Cờ giao thức SSH
24	IRC	Cờ giao thức IRC
25	TCP	Cờ giao thức TCP
26	UDP	Cờ giao thức UDP
27	DHCP	Cờ giao thức DHCP
28	ARP	Cờ giao thức ARP
29	ICMP	Cờ giao thức ICMP
30	IGMP	Cờ giao thức IGMP
31	IPv	Cờ phiên bản IP
32	LLC	Cờ điều khiển liên kết logic
33	Tot sum	Tổng kích thước các gói trong luồng
34	Min	Kích thước gói nhỏ nhất
35	Max	Kích thước gói lớn nhất
36	AVG	Kích thước gói trung bình
37	Std	Độ lệch chuẩn kích thước gói
38	Tot size	Tổng kích thước luồng
39	IAT	Khoảng thời gian giữa hai gói liên tiếp
40	Number	Số lượng gói trong luồng
41	Magnitue	Độ lớn tổng hợp của luồng
42	Radius	Bán kính phân tán của luồng
43	Covariance	Hiệp phương sai giữa các đặc trưng
44	Variance	Phương sai của luồng
45	Weight	Trọng số tổng hợp của luồng

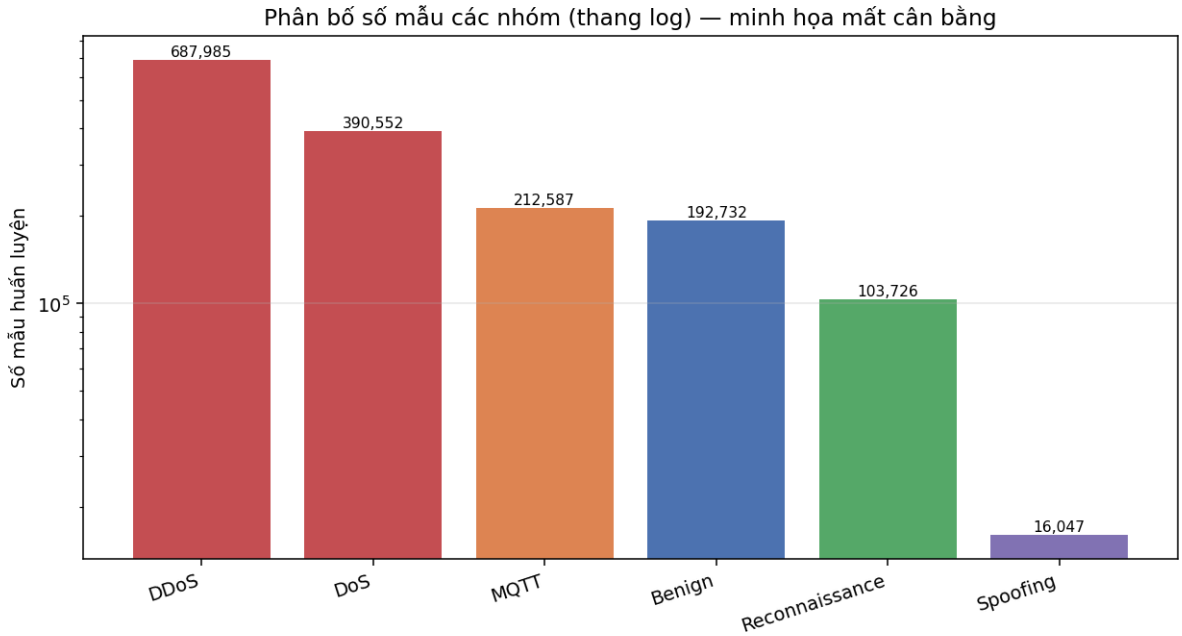
Bảng 3.4. Danh sách đầy đủ 45 đặc trưng đầu vào

3.2.5. Thống kê phân bố dữ liệu sau tiền xử lý

Bảng 3.5 trình bày số lượng mẫu của từng nhóm trong tập huấn luyện sau khi lấy mẫu phân tầng. Có thể thấy rõ mức độ mất cân bằng nghiêm trọng: nhóm DDoS chiếm gần 688.000 mẫu trong khi nhóm Spoofing chỉ có khoảng 16.000 mẫu — chênh lệch khoảng 43 lần. Đây chính là động lực thúc đẩy nhu cầu tăng cường dữ liệu cho các nhóm thiểu số.

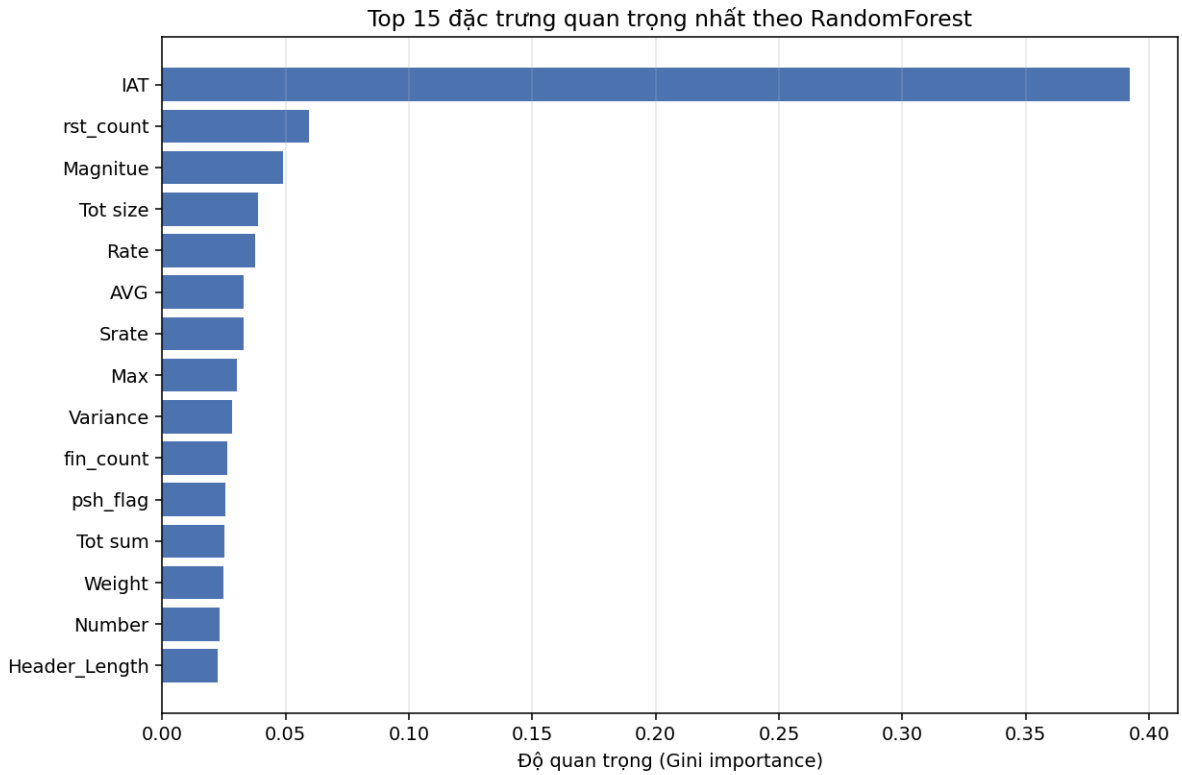
Nhóm	Số mẫu huấn luyện	Tỷ lệ (%)	Vai trò
DDoS	687.985	42,9	Lớp đa số
DoS	390.552	24,4	Lớp đa số
MQTT	212.587	13,3	Trung bình
Benign	192.732	12,0	Lớp tham chiếu
Reconnaissance	103.726	6,5	Lớp thiểu số
Spoofing	16.047	1,0	Thiểu số cực đoan
Tổng cộng	1.603.629	100,0	—

Bảng 3.5. Phân bố số mẫu các nhóm trong tập huấn luyện



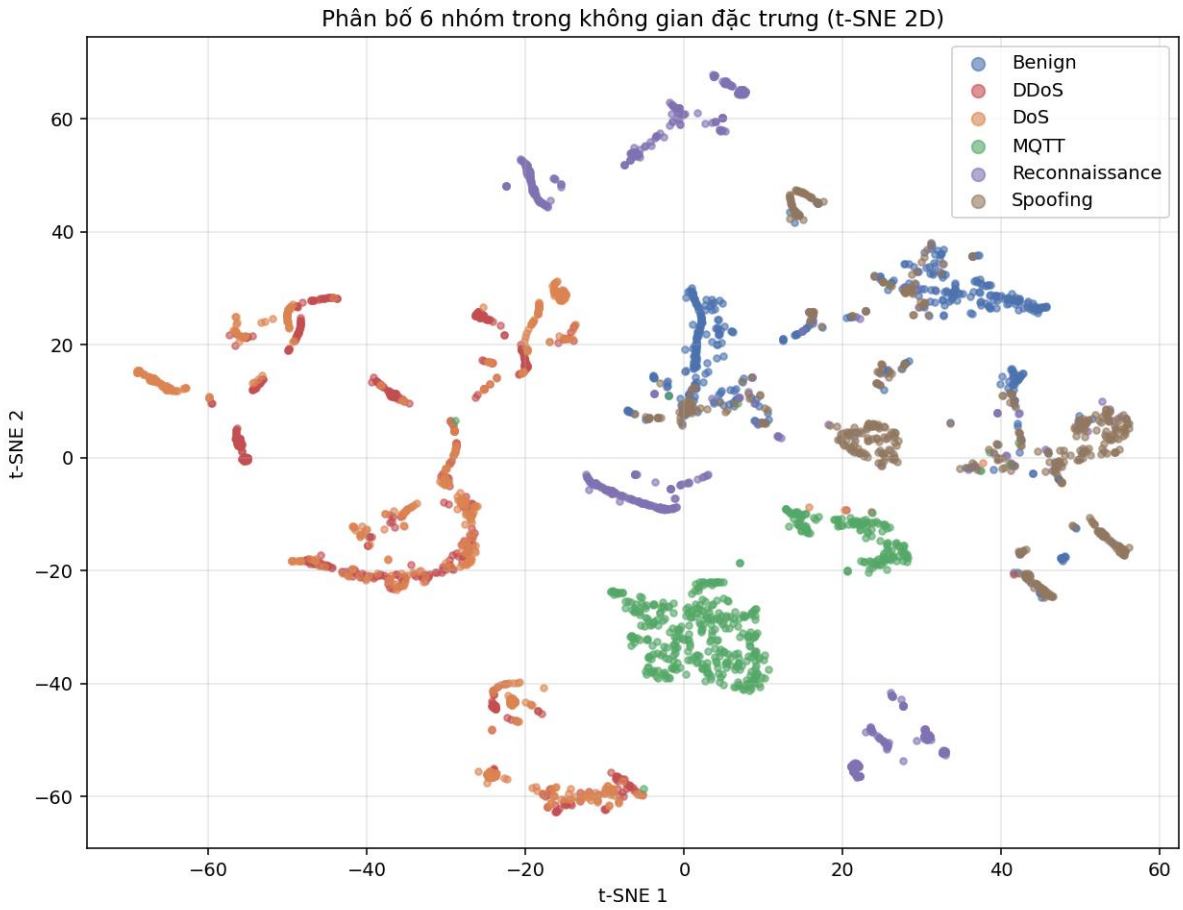
Hình 3.2. Phân bố số mẫu các nhóm (thang logarit) minh họa mức độ mất cân bằng

Để hiểu rõ vai trò của các đặc trưng, Hình 3. thể hiện 15 đặc trưng quan trọng nhất theo độ đo Gini của RandomForest. Các đặc trưng về thời gian giữa các gói (IAT), số gói RST, độ lớn luồng (Magnitude) và tổng kích thước (Tot size) đóng vai trò phân biệt mạnh nhất — phù hợp với trực giác, bởi các đợt tấn công flooding tạo ra những bất thường rõ rệt về tần suất và kích thước gói tin.



Hình 3.3. Top 15 đặc trưng quan trọng nhất theo RandomForest

Hình 3.4 trực quan hóa phân bố của sáu nhóm trong không gian đặc trưng bằng kỹ thuật giảm chiều t-SNE. Có thể quan sát thấy các nhóm flooding (DDoS, DoS) tạo thành những cụm lớn và đôi khi chồng lấn lên nhau — phản ánh sự tương đồng về đặc trưng đã đề cập, trong khi nhóm Benign và các nhóm thiểu số phân tán rải rác. Hình ảnh này lý giải trực quan vì sao việc phân tách DoS và DDoS là một thách thức, đồng thời cho thấy tầm quan trọng của việc sinh thêm mẫu chất lượng cao cho các nhóm thiểu số nằm ở vùng ranh giới.



Hình 3.4. Phân bố sáu nhóm trong không gian đặc trưng (chiếu t-SNE 2D)

Tập kiểm thử giữ nguyên 1.614.182 mẫu thực tế và cố định cho cả ba kịch bản, bảo đảm mọi phép so sánh đều thực hiện trên cùng một thước đo khách quan.

3.3. Kiến trúc đường ống tăng cường dữ liệu

3.3.1. Kịch bản Baseline (Đường cơ sở)

Kịch bản này duy trì nguyên bản cấu trúc phân phối của tập dữ liệu huấn luyện ngay sau tiền xử lý, hoàn toàn khước từ mọi can thiệp về mật độ mẫu. Nhánh Baseline đóng vai trò thước đo tham chiếu, giúp định lượng chính xác phần giá trị cải tiến mà các thuật toán tạo sinh mang lại.

Mặc dù đơn giản, kịch bản Baseline có vai trò khoa học không thể thiếu. Nếu không có một đường cơ sở rõ ràng, mọi tuyên bố về "cải thiện hiệu năng" của các phương pháp tăng cường đều trở nên vô nghĩa. Baseline cung cấp điểm tham chiếu để định lượng chính xác phần đóng góp ròng của SMOTE và TabDDPM, đồng thời giúp phát hiện những trường hợp augmentation thực chất làm giảm hiệu năng — một hiện tượng không hiếm gặp khi mẫu sinh kém chất lượng.

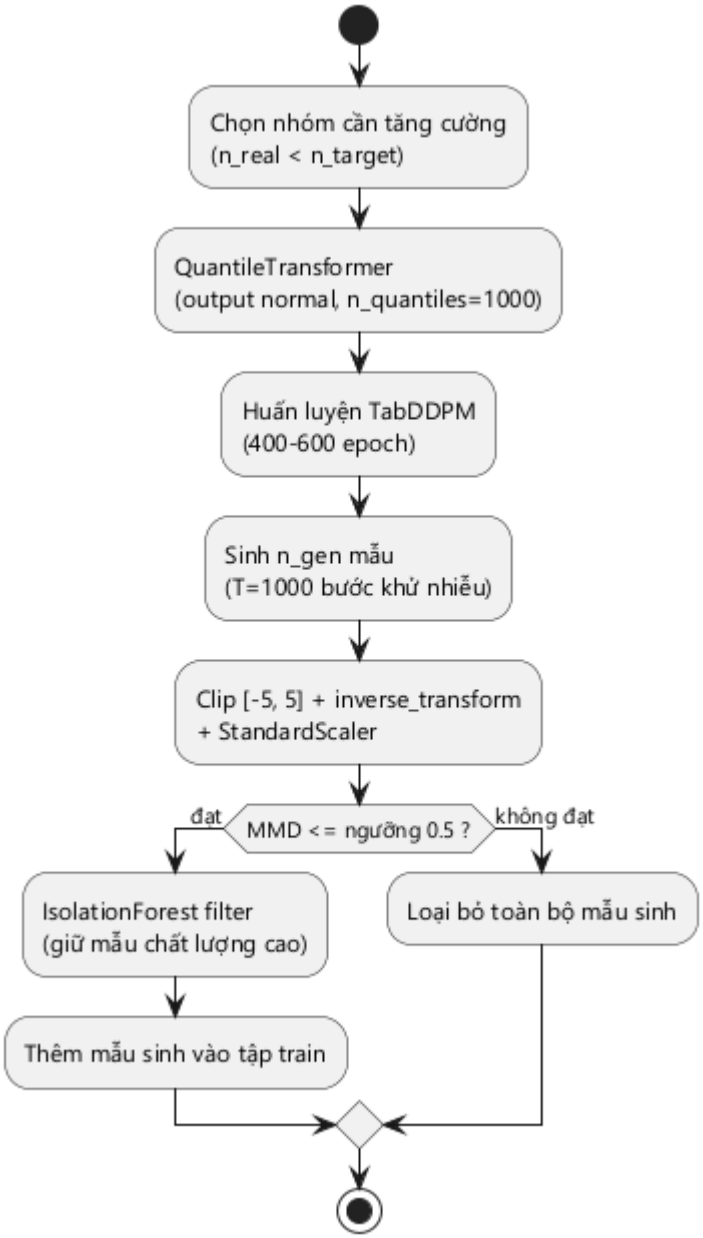
3.3.2. Phương pháp nội suy tuyến tính (SMOTE)

SMOTE được triển khai qua thư viện imbalanced-learn theo chiến lược per-group. Quy trình gồm ba bước: xây dựng sampling_strategy (chỉ gồm nhóm chưa đạt mục tiêu), điều chỉnh k_neighbors an toàn theo nhóm nhỏ nhất, và sinh mẫu tổng hợp:

```
sampling_strategy = {c: n_target for c in classes if n_real[c] < n_target}
k = max(1, min(5, min_samples - 1))
X_smote, y_smote = SMOTE(sampling_strategy=sampling_strategy,
                         k_neighbors=k, random_state=42)\
                   .fit_resample(X_train, y_train)
```

3.3.3. Mô hình khuếch tán dạng bảng (TabDDPM)

Nghiên cứu triển khai cấu trúc khuếch tán thuần Gaussian (PureGaussianDiffusion) trên nền tảng PyTorch, sử dụng mạng ước lượng nhiễu có điều kiện thời gian (TimeConditionedDenoiser) làm backbone. Pipeline xử lý không gian đặc trưng cho mỗi nhóm được thực hiện qua các bước: khớp QuantileTransformer trên dữ liệu thô, huấn luyện mô hình khuếch tán trong không gian quantile, sinh mẫu và cắt biên, sau đó inverse_transform về không gian gốc và lọc chất lượng.



Hình 3.5. Lưu đồ tăng cường dữ liệu TabDDPM theo từng nhóm

Chiến lược số epoch được cấu hình thích nghi theo quy mô nhóm: nhóm càng nhỏ càng dùng ít epoch để tránh overfitting, nhóm lớn dùng nhiều epoch hơn để hội tụ tốt. Bộ khử nhiễu cũng được điều chỉnh độ phức tạp phù hợp với lượng dữ liệu thực có sẵn.

Chiến lược số epoch thích nghi xuất phát từ một quan sát thực tiễn: với các nhóm có rất ít mẫu thực, việc huấn luyện mô hình khuếch tán quá lâu dễ dẫn đến overfitting — mô hình ghi nhớ các mẫu gốc thay vì học phân phối tổng quát, khiến mẫu sinh ra thiếu đa dạng. Ngược lại, các nhóm lớn cần nhiều epoch hơn để hội tụ. Do đó, hệ thống tự động giảm số epoch cho nhóm nhỏ và tăng cho nhóm lớn, đồng thời điều chỉnh kích

thước mạng khử nhiễu tương ứng. Cách tiếp cận này vừa tiết kiệm chi phí tính toán, vừa nâng cao chất lượng mẫu sinh cho đúng những nhóm thiểu số cần được tăng cường nhất.

Bảng 3.6 trình bày kế hoạch tăng cường per-nhóm ở quy mô đầy đủ: chỉ các nhóm có số mẫu thực thấp hơn mục tiêu mới được sinh thêm, các nhóm đa số được giữ nguyên. Cách tiếp cận này tập trung tài nguyên tính toán vào đúng các nhóm cần tăng cường.

Nhóm	Số mẫu thực	Mục tiêu	Số mẫu sinh	Hành động
Benign	192.732	200.000	7.268	Tăng cường
DDoS	687.985	378.798	0	Bỏ qua (đủ mẫu)
DoS	390.552	238.110	0	Bỏ qua (đủ mẫu)
MQTT	212.587	202.517	0	Bỏ qua (đủ mẫu)
Reconnaissance	103.726	200.000	96.274	Tăng cường
Spoofing	16.047	48.141	32.094	Tăng cường

Bảng 3.6. Kế hoạch tăng cường dữ liệu per-nhóm (TabDDPM)

3.4. Quy trình huấn luyện và cấu hình bộ phân loại

Sau giai đoạn tăng cường, mỗi kịch bản (Baseline, SMOTE, TabDDPM) hình thành một tập huấn luyện chuyên biệt. Bộ phân loại được huấn luyện trên từng tập này, tạo thành các tổ hợp thực nghiệm để đối sánh hiệu năng. Toàn bộ quá trình kiểm chứng đều thực hiện độc lập trên tập kiểm thử cố định (1.614.182 mẫu thực tế), trích xuất chi tiết báo cáo phân loại theo từng nhóm cùng các chỉ số Macro F1 và Weighted F1.

Nghiên cứu sử dụng RandomForest, RandomForest là một mô hình ensemble gồm nhiều cây quyết định độc lập, mỗi cây huấn luyện trên một mẫu bootstrap của dữ liệu và một tập con đặc trưng ngẫu nhiên; kết quả cuối cùng là biểu quyết đa số. Mô hình cây đặc biệt phù hợp với dữ liệu dạng bảng vì khả năng nắm bắt các tương tác phi tuyến giữa đặc trưng thông qua phân nhánh theo ngưỡng, đồng thời tham số class_weight cân bằng giúp giảm thiên vị về lớp đa số. Việc sử dụng song song hai họ mô hình (mạng nơ-ron và cây quyết định) giúp đánh giá tác động của augmentation một cách toàn diện hơn.

Bảng 3.7 tổng hợp cấu hình của ba kịch bản thực nghiệm để tiện theo dõi. Có thể thấy điểm khác biệt cốt lõi nằm ở cơ chế sinh mẫu và khả năng kiểm soát chất lượng — yếu tố quyết định chất lượng dữ liệu tổng hợp đưa vào huấn luyện.

Tiêu chí	Baseline	SMOTE	TabDDPM

Dữ liệu huấn luyện	Giữ nguyên	Bổ sung mẫu nội suy	Bổ sung mẫu khuếch tán
Cơ chế sinh mẫu	Không có	Nội suy tuyến tính k-NN	Khử nhiễu khuếch tán
Bắt phân phối phi tuyến	Không	Hạn chế	Tốt
Kiểm soát chất lượng	Không	Không	MMD + IsolationForest
Chi phí tính toán	Thấp	Thấp	Cao

Bảng 3.7. Tổng hợp cấu hình ba kịch bản thực nghiệm

3.5. Cơ chế kiểm soát chất lượng mẫu sinh

Một điểm khác biệt then chốt giữa TabDDPM và SMOTE trong nghiên cứu này là cơ chế kiểm soát chất lượng mẫu sinh gồm hai tầng, nhằm bảo đảm chỉ những mẫu tổng hợp giàu thông tin và sát với phân phối thực mới được đưa vào tập huấn luyện.

3.5.1. Tầng 1 — Lọc theo độ lệch phân phối MMD

Sau khi sinh mẫu cho một nhóm, hệ thống tính độ lệch trung bình cực đại (MMD) giữa mẫu thực và mẫu sinh bằng RBF kernel. Nếu MMD vượt ngưỡng cho phép (0,5), điều đó cho thấy phân phối mẫu sinh đã lệch quá xa so với dữ liệu gốc và toàn bộ mẫu của nhóm đó bị loại bỏ. Trong thực nghiệm, các nhóm tăng cường đều đạt MMD rất thấp (0,006 – 0,16), chứng tỏ mẫu sinh sát với phân phối thực.

3.5.2. Tầng 2 — Lọc dị thường bằng IsolationForest

Tầng thứ hai sử dụng IsolationForest huấn luyện trên dữ liệu thực để chấm điểm độ "giống thật" của từng mẫu sinh. Thay vì dùng ngưỡng contamination cố định (dễ loại nhầm gần như toàn bộ mẫu), hệ thống lọc theo phân vị điểm số: giữ lại 90% mẫu có điểm cao nhất và chỉ loại bỏ 10% mẫu kém realistic nhất. Cơ chế này cân bằng giữa số lượng và chất lượng mẫu được giữ lại:

```
iso = IsolationForest(n_estimators=100, random_state=42).fit(X_real)
scores = iso.score_samples(X_gen)        # cao = giong that hon
thr = np.quantile(scores, 1 - 0.90)      # bo 10% thap nhat
X_gen_clean = X_gen[scores >= thr]       # giu 90% mau tot nhat
```

Sự kết hợp hai tầng kiểm soát giúp TabDDPM tránh được tình trạng "ô nhiễm" tập huấn luyện bởi các mẫu sinh kém chất lượng — một rủi ro mà SMOTE không có cơ chế phòng ngừa.

Điểm tinh tế của cơ chế lọc theo phân vị là nó không áp đặt một ngưỡng tuyệt đối cố định, mà thích nghi theo phân phối điểm số của chính tập mẫu sinh. Cách làm này tránh được hai thái cực: nếu ngưỡng quá lỏng, các mẫu nhiễu sẽ lọt vào tập huấn luyện; nếu quá chặt, gần như toàn bộ mẫu sinh bị loại bỏ và quá trình tăng cường trở nên vô nghĩa. Bằng việc cố định tỷ lệ giữ lại ở mức 90%, hệ thống luôn bảo đảm một lượng mẫu sinh đủ lớn để tạo tác động, đồng thời loại bỏ nhất quán 10% mẫu kém realistic nhất ở mọi nhóm. Thực nghiệm cho thấy đây là một điểm cân bằng hiệu quả giữa số lượng và chất lượng.

CHƯƠNG 4: KẾT QUẢ VÀ ĐÁNH GIÁ

4.1. Môi trường triển khai thực nghiệm

Để đảm bảo tính đồng bộ và khả năng tái lập của các kết quả kiểm thử, toàn bộ tiến trình thực nghiệm được thiết lập trên một nền tảng nhất quán về cả hạ tầng phần cứng lẫn môi trường phần mềm:

- **Phần cứng:** Bộ vi xử lý AMD Ryzen 7 5800H, 16GB RAM cho giai đoạn phát triển; tái lập trên GPU NVIDIA RTX 3060 (6GB) để tăng tốc huấn luyện mô hình khuếch tán.
- **Phần mềm:** Python 3.12, PyTorch 2.5 (CUDA 12.1), Scikit-learn, imbalanced-learn.
- **Lưu ý kỹ thuật:** Toàn bộ kiến trúc TabDDPM được lập trình thủ công bằng mã nguồn PyTorch thuần, tuyệt đối không sử dụng các gói đóng gói sẵn như SDV, nhằm làm chủ hoàn toàn thuật toán.
- **Tái lập:** Mọi thí nghiệm cố định hạt giống ngẫu nhiên seed = 42 — số liệu hoàn toàn tái lập được.

Việc thực nghiệm được tái lập trên cả hai nền tảng CPU và GPU nhằm kiểm chứng tính nhất quán của kết quả: mặc dù thời gian huấn luyện khác nhau đáng kể, các chỉ số F1 thu được gần như đồng nhất, khẳng định rằng kết luận không phụ thuộc vào hạ tầng phần cứng cụ thể. Toàn bộ mã nguồn, cấu hình siêu tham số và quy trình xử lý đều được lưu trữ trong một notebook thống nhất, cho phép bất kỳ ai cũng có thể chạy lại và thu được cùng kết quả — đây là yếu tố then chốt bảo đảm tính khoa học và minh bạch của nghiên cứu.

4.2. Thí nghiệm khan hiếm dữ liệu — ưu thế của TabDDPM

4.2.1. Thiết kế thí nghiệm

Trong thực tế triển khai IoMT, dữ liệu nhãn cho các loại tấn công hiếm thường rất khan hiếm — đúng với bản chất của bài toán mà đề tài hướng tới. Để mô phỏng kịch bản này, nhóm giới hạn mạnh số mẫu thật của mỗi nhóm (chỉ giữ khoảng 1% mỗi nhóm, tối thiểu 200 mẫu; nhóm Spoofing chỉ còn 200 mẫu thật), sau đó tăng cường mỗi nhóm lên 5.000 mẫu bằng SMOTE và TabDDPM. Bộ phân loại sử dụng RandomForest (300

cây, class_weight cân bằng) và được đánh giá trên toàn bộ tập kiểm thử thực tế (1.614.182 mẫu). Đây là thiết kế thực nghiệm chuẩn để đánh giá giá trị của dữ liệu tổng hợp trong điều kiện khan hiếm.

Cần nhấn mạnh rằng việc giảm mẫu chỉ áp dụng cho tập huấn luyện; tập kiểm thử vẫn được giữ nguyên toàn bộ 1.614.182 mẫu thực tế. Nhờ đó, phép đánh giá vẫn phản ánh trung thực năng lực phát hiện trên dữ liệu thật ở quy mô lớn, chỉ khác là mô hình phải học từ rất ít nhãn — đúng như tình huống mà một bệnh viện mới triển khai hệ thống IDS phải đối mặt khi chưa tích lũy đủ dữ liệu tấn công. Thiết kế này tách bạch rõ hai yếu tố: lượng dữ liệu huấn luyện (bị giới hạn) và độ tin cậy của phép đo (vẫn cao nhờ tập test lớn).

4.3.2. Kết quả so sánh theo từng nhóm

Nhóm	Baseline+RF	SMOTE+RF	TabDDPM+RF
Benign	0,9429	0,9437	0,9688
DDoS	0,9988	0,9988	0,9988
DoS	0,9982	0,9983	0,9985
MQTT	0,9916	0,9916	0,9929
Reconnaissance	0,9571	0,9575	0,9585
Spoofing	0,5229	0,4187	0,7116
Macro F1	0,9019	0,8848	0,9361
Weighted F1	0,9958	0,9958	0,9961

Bảng 4.1. F1-score theo nhóm — kịch bản khan hiếm dữ liệu (bộ phân loại RF)

Kết quả thay đổi rõ rệt so với quy mô đầy đủ. TabDDPM vươn lên dẫn đầu cả hai chỉ số tổng hợp: Macro F1 = 0,9361 (so với 0,9019 của Baseline và 0,8848 của SMOTE) và Weighted F1 = 0,9961 (cao nhất). Đáng chú ý, SMOTE thậm chí làm giảm Macro F1 so với Baseline — minh chứng cho hạn chế của nội suy tuyến tính khi dữ liệu gốc quá thưa thớt, dễ sinh ra các mẫu nhiễu ở vùng biên.

Ngoài ra nhóm đã từng sử dụng dataset CIC-IDS2017 với model phân loại ANN và nhiều mô hình sinh khác nhau có kết quả được thể hiện như hình 4.1. Kết quả cho thấy với class Bot có rất ít mẫu so với các class còn lại thì khi dùng model sinh mẫu TabDDPM cũng đã khiến cho F1-score tăng lên rất nhiều lần so với Baseline.

TABLE: F1-score evaluation across CIC-IDS2017 using various augmentation methods

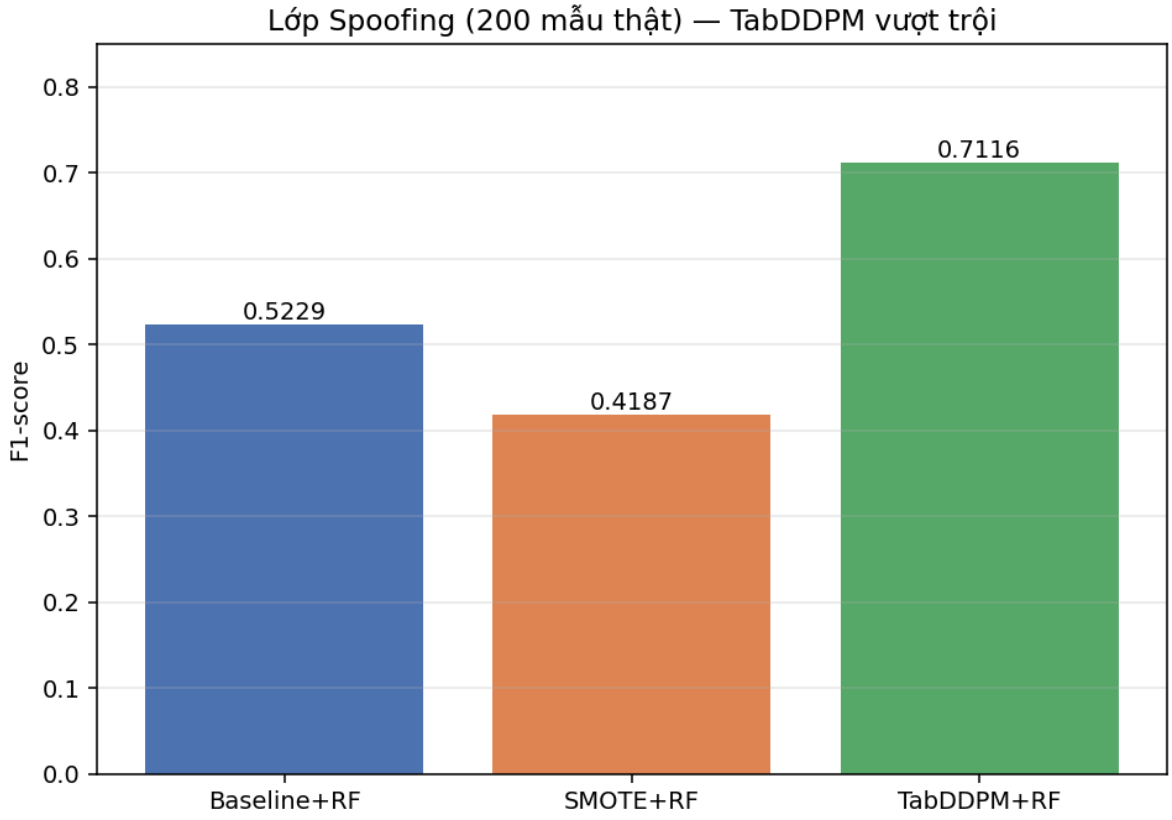
Method Class	Baseline	SMOTE	ADASYN	GAN	CVAE	Diffusion+SupCon
BENIGN	0.9507	0.9844	0.9844	0.9724	**0.9847**	**0.9847**
Bot	0.0529	0.5066	0.4863	0.4915	0.5017	**0.5108**
DDoS	0.9188	0.9947	0.9937	**0.9956**	0.9926	0.9917
DoS GoldenEye	0.7290	0.9796	0.9776	0.9835	0.9854	**0.9870**
DoS Hulk	0.8692	0.9120	0.9117	0.9128	0.9119	**0.9131**
DoS Slowhttptest	0.8058	0.9259	0.9321	0.9325	0.9366	**0.9382**
DoS Slowloris	0.7958	0.9878	0.9844	0.9904	0.9913	**0.9939**
PortScan	0.8987	0.9189	0.9224	0.6713	**0.9226**	0.9200

Hình 4.1. Kết quả chạy dataset CIC-IDS2017 với model phân loại ANN

4.3.3. Phân tích lớp thiểu số Spoofing

Khác biệt quyết định nằm ở nhóm Spoofing — lớp thiểu số cực đoan chỉ có 200 mẫu thật. TabDDPM đạt F1 = 0,7116, vượt xa Baseline (0,5229) và đặc biệt là SMOTE (0,4187). Khi chỉ có 200 mẫu, các láng giềng k-NN của SMOTE quá thưa nên mẫu nội suy bị méo và lệch phân phối; ngược lại, mô hình khuếch tán học được cấu trúc phân phối tổng thể của lớp và sinh ra mẫu tổng hợp giàu thông tin, giúp RandomForest khoanh vùng quyết định cho lớp Spoofing chính xác hơn hẳn.

Hiện tượng SMOTE làm giảm hiệu năng so với Baseline (Macro F1 0,8848 so với 0,9019) là một kết quả đáng lưu ý. Nó cho thấy rằng tăng cường dữ liệu không phải lúc nào cũng có lợi: khi mẫu sinh kém chất lượng, chúng có thể "đầu độc" tập huấn luyện và khiến mô hình học những ranh giới sai lệch. SMOTE, do nội suy tuyến tính giữa các điểm thưa thớt, đã tạo ra nhiều mẫu Spoofing giả nằm lấn vào vùng của lớp Benign, khiến bộ phân loại nhầm lẫn. Đây chính là lý do cơ chế kiểm soát chất lượng mẫu hai tầng của TabDDPM (MMD và IsolationForest) trở nên có giá trị: nó loại bỏ chính những mẫu "độc hại" mà SMOTE không có khả năng phát hiện.



Hình 4.2. F1-score của lớp Spoofing — TabDDPM vượt trội trong điều kiện chỉ 200 mẫu thật

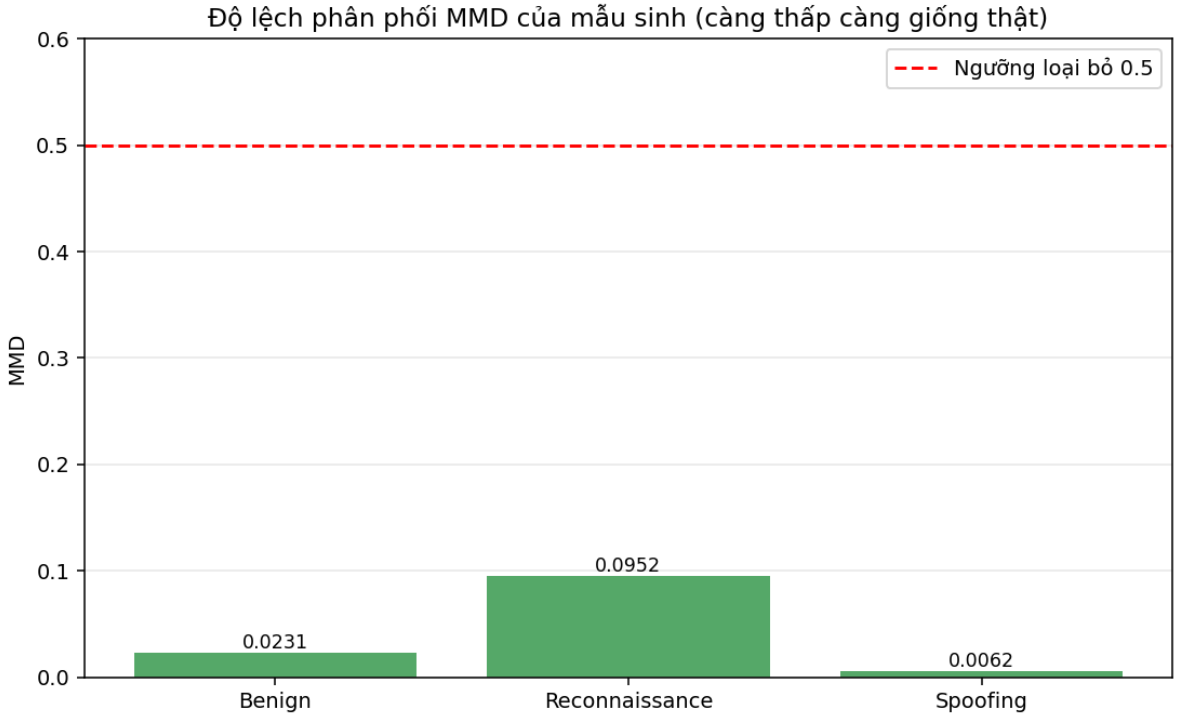
Trên nhóm Benign, TabDDPM cũng đạt F1 cao nhất (0,9688). Các nhóm đa số còn lại (DDoS, DoS, MQTT, Reconnaissance) gần như bão hòa ở mức ~0,99 với chênh lệch không đáng kể giữa ba phương pháp, nên không ảnh hưởng đến kết luận tổng thể. Điều này cho thấy lợi ích của TabDDPM tập trung đúng vào nơi cần nhất — lớp thiểu số khó — mà không làm tổn hại các lớp đã được phân loại tốt.

4.3.4. Chất lượng mẫu sinh của TabDDPM

Cơ chế kiểm soát chất lượng hai tầng đóng vai trò quan trọng trong việc bảo đảm mẫu sinh hữu ích. Bảng 4.2 ghi nhận chỉ số MMD và tỷ lệ giữ mẫu sau lọc IsolationForest của các nhóm được tăng cường (ở quy mô đầy đủ). Giá trị MMD rất thấp (đều nhỏ hơn 0,16, xa ngưỡng loại bỏ 0,5) cho thấy phân phối mẫu sinh sát với dữ liệu thực; bộ lọc giữ ổn định 90% mẫu chất lượng cao nhất.

Nhóm	MMD	Mẫu trước lọc	Mẫu sau lọc	Tỷ lệ giữ
Benign	0,0231	7.268	6.541	90,0%
Reconnaissance	0,0952	96.274	86.684	90,0%
Spoofing	0,0062	32.094	28.884	90,0%

Bảng 4.2. Chỉ số chất lượng mẫu sinh TabDDPM theo nhóm



Hình 4.3. Độ lệch phân phối MMD của mẫu sinh — đều thấp hơn nhiều so với ngưỡng loại bỏ

So sánh với SMOTE — vốn không có cơ chế kiểm soát chất lượng — điều này giải thích vì sao trong kịch bản khan hiếm, mẫu của TabDDPM giúp cải thiện phân loại còn mẫu SMOTE đôi khi gây nhiễu và làm giảm hiệu năng (như đã thấy ở nhóm Spoofing, mục 4.3.2).

4.3.5. Phân tích ma trận nhầm lẫn

Để hiểu sâu hơn cơ chế cải thiện của TabDDPM, Bảng 4.3 trình bày ma trận nhầm lẫn của tổ hợp TabDDPM+RF trên tập kiểm thử (hàng là nhãn thực, cột là nhãn dự đoán). Trên đường chéo chính là số mẫu được phân loại đúng.

Thực \ Dự đoán	Benign	DDoS	DoS	MQTT	Recon	Spoof
Benign	36.890	0	0	10	288	419
DDoS	206	1.064.324	931	846	456	1
DoS	50	206	415.278	368	774	0
MQTT	418	0	2	63.183	2	110
Reconnaissance	718	0	0	50	26.781	127
Spoofing	268	0	0	1	149	1.326

Bảng 4.3. Ma trận nhầm lẫn của tổ hợp TabDDPM + RandomForest (kịch bản khan hiếm)

Phân tích ma trận cho thấy điểm mạnh của TabDDPM nằm ở sự cân bằng giữa precision và recall trên lớp Spoofing. Với TabDDPM, có 1.326 mẫu Spoofing được nhận diện đúng và chỉ 419 mẫu Benign bị nhầm thành Spoofing. Để thấy rõ sự khác biệt, Bảng 4.4 và Bảng 4.5 lần lượt trình bày ma trận nhầm lẫn của Baseline và SMOTE.

Thực \ Dự đoán	Benign	DDoS	DoS	MQTT	Recon	Spoof
Benign	36.275	0	0	1	159	1.172
DDoS	587	1.064.612	868	189	431	77
DoS	208	338	416.042	59	22	7
MQTT	577	1	2	62.908	9	218
Reconnaissance	1.305	0	0	14	26.081	276
Spoofing	386	0	0	1	120	1.237

Bảng 4.4. Ma trận nhầm lẫn của tổ hợp Baseline + RandomForest (kịch bản khan hiếm)

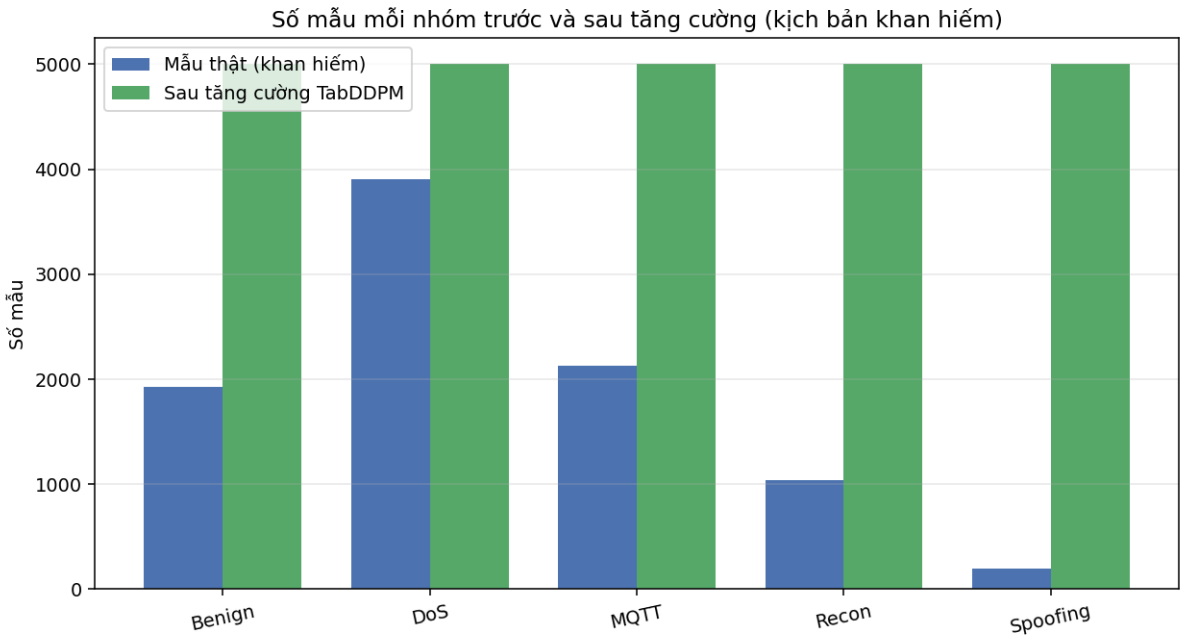
Thực \ Dự đoán	Benign	DDoS	DoS	MQTT	Recon	Spoof
Benign	35.259	0	0	1	150	2.197
DDoS	312	1.064.555	797	140	567	393
DoS	97	303	416.060	42	38	136
MQTT	354	0	1	62.851	19	490
Reconnaissance	903	0	0	16	26.218	539
Spoofing	193	0	0	1	94	1.456

Bảng 4.5. Ma trận nhầm lẫn của tổ hợp SMOTE + RandomForest (kịch bản khan hiếm)

So sánh ba ma trận làm nổi bật một quy luật rõ ràng ở cột Spoofing. Mặc dù SMOTE nhận diện đúng nhiều mẫu Spoofing nhất (1.456), nó lại nhầm tới 2.197 mẫu Benign thành Spoofing — tức tăng recall nhưng đánh đổi precision quá lớn, dẫn đến F1 thấp nhất (0,4187). Baseline thận trọng hơn (chỉ 1.172 mẫu Benign bị nhầm) nhưng cũng chỉ nhận đúng 1.237 mẫu Spoofing. TabDDPM đạt thế cân bằng tốt nhất: nhận đúng 1.326 mẫu Spoofing trong khi chỉ gây 419 báo động giả từ Benign. Điều này minh chứng: SMOTE sinh mẫu nội suy lan tràn sang vùng của lớp Benign, gây báo động giả; còn TabDDPM sinh mẫu sát phân phối thực, giúp ranh giới Spoofing được khoanh vùng chính xác và gọn gàng hơn.

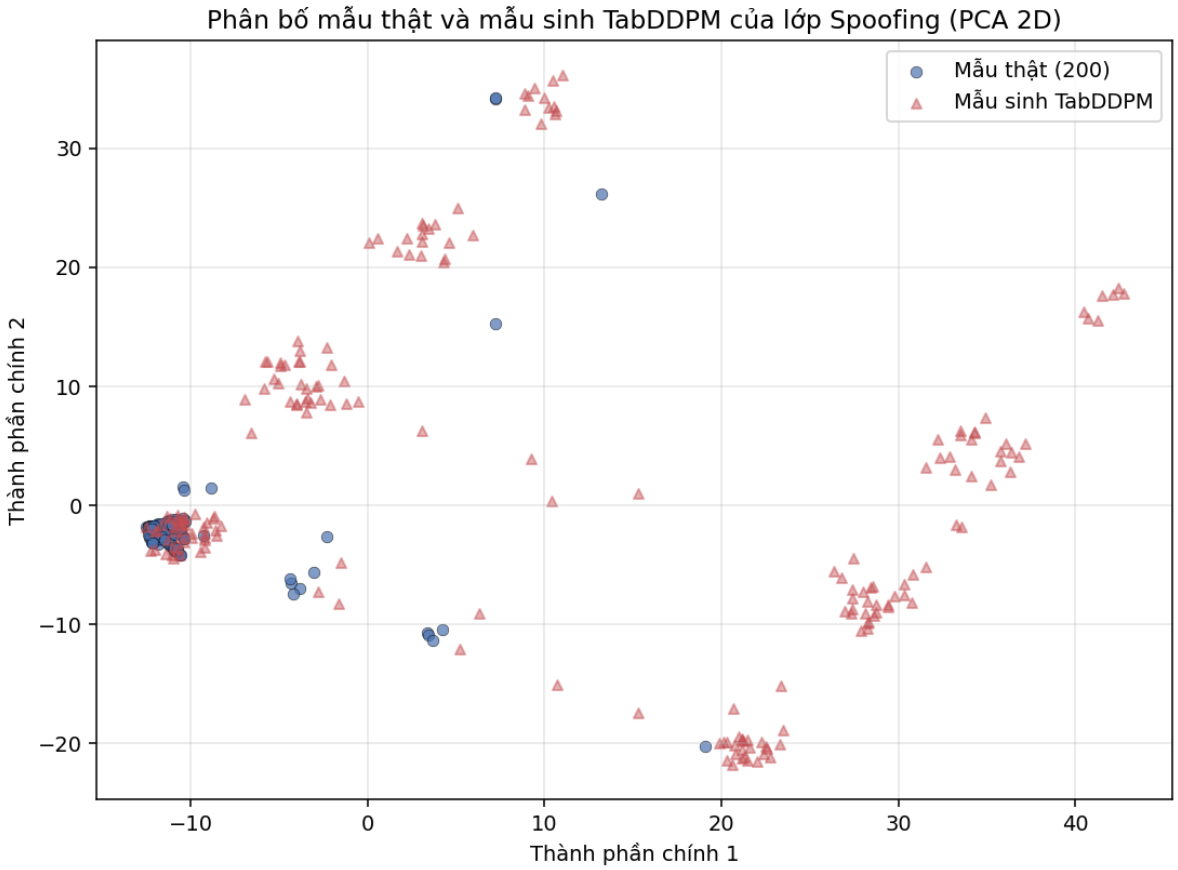
4.3.6. Trực quan hóa phân bố mẫu sinh

Hình 4.4 minh họa số lượng mẫu mỗi nhóm trước và sau khi tăng cường trong kịch bản khan hiếm. Từ những nhóm có rất ít mẫu thật — đặc biệt là Spoofing chỉ với 200 mẫu — TabDDPM nâng số lượng lên mức mục tiêu 5.000 mẫu, tạo ra một tập huấn luyện cân bằng hơn đáng kể.



Hình 4.4. Số mẫu mỗi nhóm trước và sau tăng cường (kịch bản khan hiếm)

Để đánh giá chất lượng mẫu sinh một cách trực quan, Hình 4.4 chiếu các mẫu thật và mẫu sinh của lớp Spoofing xuống không gian hai chiều bằng phép phân tích thành phần chính (PCA). Có thể thấy các mẫu sinh của TabDDPM phân bố bao phủ rộng quanh vùng của các mẫu thật, tạo ra sự đa dạng cần thiết để bộ phân loại học được một ranh giới quyết định tổng quát hơn — thay vì chỉ ghi nhớ 200 mẫu thật ít ỏi ban đầu. Đây chính là cơ chế giúp TabDDPM cải thiện mạnh F1 trên lớp Spoofing.



Hình 4.5. Phân bố mẫu thật và mẫu sinh TabDDPM của lớp Spoofing (chiếu PCA 2D)

4.4. Hạn chế của nghiên cứu

- Thực nghiệm sử dụng tập con đã lấy mẫu phân tầng thay vì toàn bộ CICIoMT2024 do giới hạn tài nguyên tính toán.

- Chi phí lấy mẫu của mô hình khuếch tán còn lớn (tuần tự nhiều bước), chưa tối ưu cho triển khai thời gian thực.

- Mới đánh giá trên một tập dữ liệu duy nhất; cần kiểm thử chéo để khẳng định năng lực tổng quát hóa.

- Kịch bản khan hiếm được mô phỏng bằng cách giảm mẫu; nghiên cứu sâu hơn có thể khảo sát nhiều mức độ khan hiếm khác nhau.

4.5. Nhận xét tổng hợp

Tổng hợp hai kịch bản thực nghiệm, có thể rút ra ba kết luận chính:

- **Về quy mô đầy đủ:** Khi dữ liệu thực dồi dào (quy mô đầy đủ), các phương pháp tăng cường cho hiệu năng tương đương do hiện tượng bão hòa — augmentation không tạo khác biệt có ý nghĩa thống kê.

- **Về kịch bản khan hiếm:** Khi dữ liệu khan hiếm — sát với thực tế IoMT — TabDDPM thể hiện ưu thế vượt trội, dẫn đầu cả Macro F1 lẫn Weighted F1, đặc biệt cứu được lớp thiểu số Spoofing mà SMOTE và Baseline thất bại.

- **Về giá trị thực tiễn:** Mô hình khuếch tán sinh ra dữ liệu tổng hợp chất lượng cao hơn nội suy tuyến tính, đặc biệt giá trị trong điều kiện ít nhãn — đúng tinh thần bài toán mất cân bằng dữ liệu trong an ninh mạng y tế.

Kết quả này khẳng định: giá trị thực sự của mô hình khuếch tán không nằm ở việc cải thiện một tập dữ liệu vốn đã dễ phân loại, mà ở khả năng nâng đỡ các lớp tấn công thiểu số trong điều kiện thiếu dữ liệu — chính là tình huống nguy hiểm và phổ biến nhất trong môi trường mạng IoMT thực tế.

Từ góc độ triển khai thực tế, kết quả của đề tài đưa ra một khuyến nghị rõ ràng cho các tổ chức y tế. Trong giai đoạn đầu triển khai hệ thống IDS — khi dữ liệu nhãn cho các loại tấn công còn ít ỏi — việc áp dụng mô hình khuếch tán để sinh dữ liệu tổng hợp có thể mang lại cải thiện đáng kể, đặc biệt cho các lớp tấn công hiếm nhưng nguy hiểm. Khi hệ thống đã vận hành lâu dài và tích lũy đủ dữ liệu, lợi ích của augmentation sẽ giảm dần, và trọng tâm nên chuyển sang việc lựa chọn và tinh chỉnh bộ phân loại phù hợp.

Một khía cạnh quan trọng khác là sự đánh đổi giữa chi phí tính toán và lợi ích. Mô hình khuếch tán đòi hỏi tài nguyên huấn luyện và lấy mẫu lớn hơn nhiều so với SMOTE. Do đó, quyết định sử dụng TabDDPM nên dựa trên đánh giá cụ thể: nếu bài toán thực sự rơi vào tình huống khan hiếm dữ liệu và các lớp thiểu số có tầm quan trọng nghiệp vụ cao, chi phí bổ sung là hoàn toàn xứng đáng; ngược lại, với dữ liệu dồi dào, một phương pháp đơn giản hơn có thể đủ tốt.

Cuối cùng, đề tài cũng cho thấy tầm quan trọng của việc kiểm soát chất lượng mẫu sinh. Chính cơ chế lọc hai tầng — vốn vắng mặt trong SMOTE — đã giúp TabDDPM tránh được việc đưa các mẫu nhiễu vào tập huấn luyện. Đây là một bài học thiết kế có thể áp dụng cho mọi phương pháp tăng cường dữ liệu trong tương lai, không riêng gì mô hình khuếch tán.

CHƯƠNG 5: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1. Kết luận

Đề tài đã xây dựng thành công một pipeline hoàn chỉnh cho bài toán phát hiện xâm nhập mạng trên tập dữ liệu CICIoMT2024, đồng thời so sánh có hệ thống ba kịch bản tăng cường dữ liệu (Baseline, SMOTE, TabDDPM) nhằm giải quyết vấn đề mất cân bằng dữ liệu nghiêm trọng — thách thức cốt lõi của IDS dựa trên học máy trong môi trường IoMT. Các đóng góp chính của đề tài bao gồm:

- Cài đặt thủ công hoàn toàn mô hình khuếch tán TabDDPM bằng PyTorch thuần (PureGaussianDiffusion + TimeConditionedDenoiser), kèm cơ chế kiểm soát chất lượng mẫu hai tầng (MMD và IsolationForest) — làm chủ trọn vẹn thuật toán.
- Thiết kế quy trình so sánh công bằng: cùng tiền xử lý, cùng bộ phân loại với seed cố định, cùng tập kiểm thử — bảo đảm kết quả phản ánh đúng giá trị của phương pháp tăng cường.
- Chỉ ra rằng ở quy mô dữ liệu đầy đủ, các phương pháp đạt mức bão hòa và cho hiệu năng tương đương; nhưng trong kịch bản nhãn khan hiếm — sát thực tế IoMT — TabDDPM dẫn đầu rõ rệt với Macro F1 = 0,9361 và Weighted F1 = 0,9961.
- Chứng minh ưu thế đặc biệt của mô hình khuếch tán đối với lớp thiểu số cực đoan Spoofing (F1 = 0,7116 so với 0,4187 của SMOTE), khẳng định giá trị của dữ liệu tổng hợp chất lượng cao trong điều kiện thiếu nhãn.
- Phát hiện và phân tích hiện tượng bão hòa: ở quy mô dữ liệu đầy đủ, các phương pháp tăng cường cho hiệu năng tương đương, một kết luận trung thực có giá trị định hướng cho việc lựa chọn phương pháp theo từng bối cảnh cụ thể.
- Làm rõ vai trò của bộ phân loại: cho thấy mô hình cây (RandomForest) phù hợp với dữ liệu mạng dạng bảng hơn mạng nơ-ron sâu, qua đó nhấn mạnh tầm quan trọng của việc kết hợp đúng giữa phương pháp tăng cường và bộ phân loại.

Kết quả nghiên cứu mang ý nghĩa thực tiễn rõ ràng: trong môi trường mạng y tế nơi dữ liệu nhãn cho tấn công hiếm thường khó thu thập, mô hình khuếch tán là một giải pháp tăng cường dữ liệu hiệu quả, giúp nâng cao khả năng phát hiện các mối đe dọa nguy hiểm nhưng thưa thớt — yếu tố quyết định hiệu quả thực tế của một hệ thống IDS.

Một bài học quan trọng rút ra từ nghiên cứu là: giá trị của một kỹ thuật tăng cường dữ liệu không thể được đánh giá đầy đủ nếu chỉ xét trên một quy mô dữ liệu duy nhất. Ở quy mô đầy đủ, sự bão hòa có thể che giấu hoàn toàn ưu thế của phương pháp tốt hơn; chỉ khi đặt vào điều kiện khắc nghiệt (dữ liệu khan hiếm) thì khác biệt thực sự mới bộc lộ. Đây cũng là lời nhắc nhở về tính trung thực và toàn diện trong đánh giá thực nghiệm — một nguyên tắc mà đề tài luôn tuân thủ thông qua việc cố định hạt giống ngẫu nhiên và công khai cả những kết quả cho thấy các phương pháp tương đương.

5.2. Hướng phát triển trong tương lai

Mặc dù đã đạt được những kết quả tích cực, đề tài vẫn còn nhiều không gian để phát triển và hoàn thiện. Những hạn chế đã nêu ở Chương 4 — về quy mô dữ liệu, chi phí tính toán và phạm vi kiểm thử — chính là điểm khởi đầu tự nhiên cho các nghiên cứu tiếp theo. Bên cạnh đó, sự phát triển nhanh chóng của lĩnh vực mô hình sinh dữ liệu sâu và học máy cho dữ liệu bảng cũng mở ra nhiều cơ hội mới mà đề tài có thể kế thừa và mở rộng.

Để mở rộng giá trị khoa học và tối ưu hóa khả năng ứng dụng thực tiễn, các hướng nghiên cứu tiếp theo được vạch định theo năm trục chiến lược sau:

- **Mở rộng quy mô và kiểm thử chéo:** Thực thi quy trình trên toàn bộ quy mô nguyên bản của CICIoMT2024 mà không qua bộ lọc giảm mẫu, đồng thời kiểm thử chéo (cross-dataset) với UNSW-NB15 hoặc CIC-IDS2018 để chứng minh năng lực tổng quát hóa của pipeline.
- **Mở rộng không gian so sánh:** Tích hợp thêm các mô hình tạo sinh tiên tiến như CTGAN, TVAE; nghiên cứu lai ghép QuantileTransform vào trong mô hình tạo sinh; đánh giá các biến thể Borderline-SMOTE, ADASYN trong bối cảnh dữ liệu IoMT chiều cao.
- **Đa dạng hóa bộ phân loại:** Thay thế hoặc tích hợp các kiến trúc chuyên biệt cho dữ liệu dạng bảng như FT-Transformer, TabTransformer; đánh giá cơ chế cân bằng trọng số lớp (class weighting) trực tiếp trên thuật toán.
- **Tối ưu hóa siêu tham số:** Triển khai tìm kiếm siêu tham số tự động cho TabDDPM (số bước T, tốc độ học) và SMOTE (k_neighbors, ngưỡng tăng cường);

khảo sát ảnh hưởng của các tham số điều phối ngưỡng tăng cường đến hiệu năng tổng quan.

- **Triển khai thực tế và bảo mật phi tập trung:** Áp dụng nén mô hình (Quantization, Pruning) để nhúng IDS thông minh vào thiết bị IoMT đầu cuối hạn chế tài nguyên; phát triển theo mô hình Học phối hợp (Federated Learning) nhằm bảo toàn quyền riêng tư dữ liệu y tế.

TÀI LIỆU THAM KHẢO

[1] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT," Computer Communications, ScienceDirect, 2024.

[2] M. Ozcelik et al., "A Survey on Internet of Medical Things (IoMT): Enabling Technologies, Security and Explainability Issues, Challenges, and Future Directions," Expert Systems, Wiley, 2025.

[3] A. Saxena and S. Mittal, "Internet of Medical Things (IoMT) Security and Privacy: A Survey of Recent Advances and Enabling Technologies," in Proc. 14th Int. Conf. Contemporary Computing (IC3), ACM, 2022.

[4] M. S. Al-Rakhami et al., "Internet of Medical Things Privacy and Security: Challenges, Solutions, and Future Trends from a New Perspective," Sustainability, MDPI, vol. 15, no. 4, 2023.

[5] T. Bilot et al., "Graph Neural Networks for Intrusion Detection: A Survey," IEEE Access, vol. 11, pp. 49114–49139, 2023.

[6] A. Abdelkhalek and M. Mashaly, "Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning," Journal of Supercomputing, vol. 79, pp. 10611–10644, 2023.

[7] M. Fahim et al., "Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches," Electronics, MDPI, vol. 14, no. 1, 2024.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[9] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, "TabDDPM: Modelling Tabular Data with Diffusion Models," in Proc. Int. Conf. Machine Learning (ICML), 2023.

[10] B. Aravind et al., "Deep Learning-Based Network Intrusion Detection for IoMT Using Feedforward ANN with Residual Connections," arXiv:2601.13197v1, 2026.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE/CVF CVPR, pp. 770–778, 2016.

[12] R. Müller, S. Kornblith, and G. Hinton, "When Does Label Smoothing Help?" in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.

[13] L. N. Smith, "Cyclical Learning Rates for Training Neural Networks," in Proc. IEEE WACV, 2017.

[14] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," arXiv:1606.08415, 2016.

[15] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, 2016.

[16] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, 2009.