THIẾT KẾ MÔ HÌNH AUTO-ENCODER ĐỂ PHÁT HIỆN DỊ THƯỜNG ÂM THANH

DESIGN OF AUTO-ENCODER MODEL FOR AUDIO ANONYMOUS DETECTION

Nguyễn Thiên Tân[1], Đỗ Quang Hiệp[2]

[1]Khoa Điện tử và Kỹ thuật máy tính, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

[2]Khoa Điện - Tự động hóa, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp

Đến Tòa soạn ngày 15/7/2025, chấp nhận đăng ngày 10/08/2025

Tóm tắt: Trong bài báo này, chúng tôi đề xuất một hệ thống phát hiện bất thường âm thanh phi giám sát dựa trên các mô hình học sâu và học máy truyền thống, bao gồm LSTM Autoencoder, GRU Autoencoder và One-Class SVM. Dữ liệu đầu vào là log-Mel spectrogram được chuẩn hóa từ các mẫu âm thanh môi trường đô thị. Hệ thống không yêu cầu nhãn và sử dụng lỗi tái tạo làm tiêu chí chính để phát hiện bất thường. Kết quả thực nghiệm cho thấy LSTM Autoencoder đạt hiệu suất cao nhất với MAE = 0.0172, RMSE = 0.0334 và MAPE = 2.84%, vượt trội so với GRU và OCSVM. Biểu đồ t-SNE xác nhận rằng các mẫu bất thường được ánh xạ tách biệt rõ rệt trong không gian đặc trưng. Bên cạnh đó, hệ thống có tiềm năng cao để triển khai trên các thiết bị biên (edge) hoặc trong hệ thống IoT nhằm giám sát âm thanh thời gian thực mà không cần giám sát.

Từ khóa: Phát hiện bất thường, Autoencoder, LSTM, GRU, OCSVM, log-Mel spectrogram, âm thanh, học không giám sát, IoT.

Abstract: This paper proposes an unsupervised acoustic anomaly detection system utilizing both deep learning and conventional machine learning models, including LSTM Autoencoder, GRU Autoencoder, and One-Class SVM. The input data consist of normalized log-Mel spectrograms derived from diverse urban environmental sounds. The system does not require labeled data and relies on reconstruction errors to detect anomalies. Experimental results demonstrate that the LSTM Autoencoder outperforms other models with MAE = 0.0172, RMSE = 0.0334, and MAPE = 2.84%. A 3D t-SNE visualization confirms that the anomalous samples are distinctly separated in the learned feature space. Moreover, the system holds strong potential for deployment in edge devices or IoT-based real-time acoustic monitoring applications without supervision.

Keywords: Anomaly detection, Autoencoder, LSTM, GRU, OCSVM, log-Mel spectrogram, audio, unsupervised learning, IoT.

1. GIỚI THIỆU

Với sự phát triển nhanh chóng của thành phố thông minh, nhà thông minh và giám sát công nghiệp, nhu cầu phát hiện bất thường trong môi trường âm thanh ngày càng cấp thiết. Các tình huống như la hét, đập phá, còi báo cháy hay va chạm mạnh có thể là dấu hiệu cảnh báo tội phạm, tai nạn, cháy nổ hoặc sự cố thiết bị. Tuy nhiên, việc gán nhãn dữ liệu cho hệ thống phát hiện bất thường vừa tốn kém, vừa mất thời gian và dễ thiếu sót. Đặc biệt, các tình huống bất thường hiếm khi xảy ra, dẫn đến dữ liệu mất cân bằng, làm giảm hiệu quả của các mô hình học có giám sát [1], [2]. Do đó, học không giám sát và học bán giám sát trở thành hướng tiếp cận tiềm năng, trong đó mô hình học đặc trưng của hành vi bình thường và coi

sai lệch trong tái tạo hoặc phân phối là dấu hiệu bất thường [3].

Các nghiên cứu gần đây cho thấy Autoencoder, đặc biệt là LSTM Autoencoder và GRU Autoencoder, đạt hiệu quả cao với chuỗi tín hiệu như cảm biến, video và âm thanh [4], [5]. Các mạng này có khả năng học biểu diễn ẩn từ dữ liệu chuỗi và phát hiện bất thường thông qua sai số tái tạo cao. Bên cạnh đó, One-Class SVM (OCSVM) vẫn là phương pháp kinh điển hiệu quả để xác định ranh giới dữ liệu bình thường trong không gian vector hóa.

Gần đây, nhiều công trình tập trung khai thác mô hình học sâu không giám sát cho bài toán này. Nhóm Marchi và cộng sự là những người tiên phong áp dụng LSTM Autoencoder cho dữ liệu âm thanh công nghiệp, chứng minh khả năng học mẫu âm thanh bình thường và phát hiện bất thường qua lỗi tái tạo [16]. Tại DCASE Challenge, Koizumi và cộng sự triển khai pipeline phát hiện bất thường với CNN và AE kết hợp hậu xử lý, song phần lớn hệ thống vẫn cần nhãn bất thường rõ ràng hoặc thiết kế chuyên biệt cho từng thiết bị [10]. Gemmeke [11] và Salamon [12] đã xây dựng các bộ dữ liệu nền như ESC-50 và UrbanSound8K, hỗ trợ nhận dạng âm thanh đô thị, nhưng ít được khai thác trong bối cảnh phát hiện bất thường không giám sát. Nghiên cứu [13] còn kết hợp OCSVM với học sâu để tận dụng biểu diễn phi tuyến và phân tách tuyến tính.

Bố cục bài báo gồm phần 1 giới thiệu chung, phần 2 trình bày chi tiết về phương pháp xử lý log-Mel, kiến trúc mạng và phương pháp phát hiện bất thường, phần 3 mô tả thí nghiệm, đánh giá định lượng, trực quan hóa kết quả và phần 4 là kết luận vấn đề nghiên cứu.

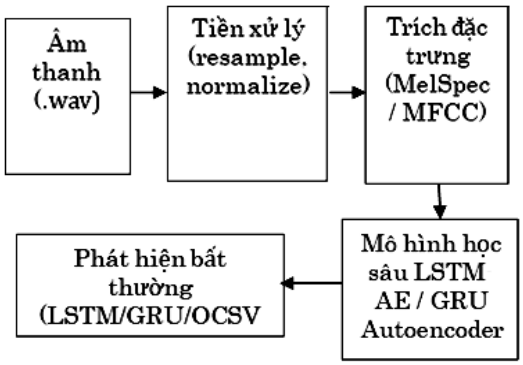
2. PHƯƠNG PHÁP ĐỀ XUẤT

Chúng tôi đề xuất phương pháp phát hiện bất thường âm thanh lai, kết hợp ba mô hình độc lập nhưng bổ sung cho nhau: LSTM Autoencoder, GRU Autoencoder và One-Class SVM (OCSVM). Cách tiếp cận này sử dụng các mô hình học sâu tuần tự để học đặc trưng âm thanh theo thời gian và một mô hình học máy cổ điển để phân vùng dữ liệu trong không gian đặc trưng.

2.1. Kiến trúc hệ thống

Đầu vào của hệ thống là các đoạn âm thanh (clip .wav) được chuyển đổi sang log-Mel spectrogram, một dạng đặc trưng phổ biến trong các bài toán nhận dạng âm thanh [1], [9]. Đặc trưng này mô phỏng cách con người cảm nhận âm thanh qua thang Mel giúp giảm nhiễu và tăng tính phân biệt giữa các loại âm thanh.

Hệ thống phát hiện bất thường âm thanh được đề xuất gồm 4 giai đoạn chính.



Hình 1. Sơ đồ tổng quan của hệ thống

2.2. Biểu diễn đầu vào: Log-Mel Spectrogram

Dữ liệu thu thập từ IDMT-Traffic[14], một dataset dùng trong nghiên cứu giám sát giao thông bằng âm thanh (ATM - Acoustic Traffic Monitoring).

Mục tiêu là tạo điều kiện phát triển các thuật toán nhận dạng âm thanh cho các thiết bị

nhúng có giới hạn về chất lượng mic và năng lực tính toán. Đặc điểm của dataset: Gồm 2.5 giờ âm thanh stereo ghi lại 4718 sự kiện xe đi qua. Dữ liệu được thu từ cả micro cao cấp (sE8) và micro MEMS chất lượng thấp hơn. Ghi âm tại 4 địa điểm với các điều kiện vận tốc và môi trường khác nhau (đường khô/ướt, 30/50/70 km/h). Gồm 4 lớp phương tiện: car, truck, motorcycle, bus. Dữ liệu âm thanh đầu vào được xử lý thành log-Mel spectrogram với các tham số:

Bảng 1. Dữ liệu tham chiếu

Thông số	Giá trị
Tần số mẫu (sample rate) Hz	22050
FFT window size	2048
Hop length	512
Số băng tần Mel	128
Thời lượng chuẩn hóa mỗi đoạn (giây)	~3
Số khung thời gian (Time frames)	256

Spectrogram được chuyển đổi qua thang logarit để giảm ảnh hưởng biên độ:

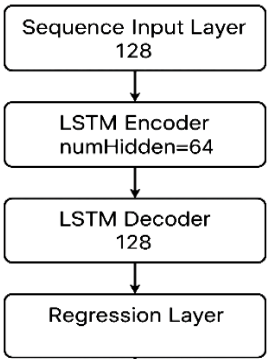
$$S_{\log}(t, f) = \log\left(1 + \frac{|X(t, f)|}{\varepsilon}\right) \quad (1)$$

Với X(t,f) là phổ tại thời điểm t, tần số f, và $\epsilon \approx 1e{-}6$.

Tất cả các spectrogram đều được resize về kích thước [256 × 128] (Time × Frequency) và được lưu trong biến cell X{i}.

2.3. Xây dựng mô hình

Mô hình Autoencoder là mô hình học không giám sát với mục tiêu nén dữ liệu đầu vào và tái tạo lại đầu ra sao cho gần giống nhất với đầu vào ban đầu [7]. Trong đó, LSTM (Long Short-Term Memory) là một kiến trúc mạng nơron hồi tiếp (RNN) có khả năng ghi nhớ dài hạn, giúp xử lý dữ liệu chuỗi âm thanh có tính phụ thuộc thời gian [1], [3].



Hình 2. Cấu trúc LSTM Autoencoder

Dữ liệu bất thường thường có sai số tái tạo lớn do khác biệt với phân phối dữ liệu bình thường đã học được [6]. Mô hình LSTM Autoencoder gồm 2 tầng chính: Encoder: nén dãy log-Mel thành biểu diễn rút gọn (latent vector) với 64 chiều và Decoder: tái tạo lại dãy log-Mel từ biểu diễn đã mã hóa. Hàm lỗi: Mean Squared Error (MSE).

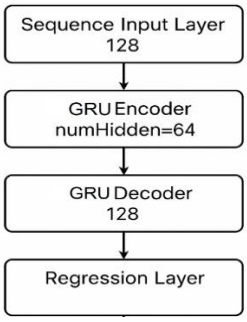
$$\text{Anomaly}_i = \begin{Bmatrix} 1, \ldots If\ errori > \theta \\ 0, \ldots If\ errori < \theta \end{Bmatrix} \quad (2)$$

Ngưỡng phát hiện bất thường, sử dụng ngưỡng thống kê 2σ:

$$\theta = \mu_{\text{error}} + 2\sigma_{\text{error}} \quad (3)$$

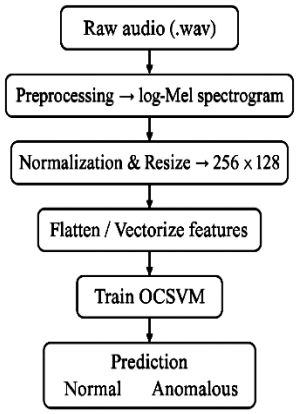
Mô hình GRU (Gated Recurrent Unit) là một biến thể đơn giản hơn của LSTM, kiến trúc GRU Autoencoder tương tự LSTM Autoencoder, chỉ thay lớp lstmLayer bằng gruLayer. GRU là biến thể đơn giản hơn của LSTM với ít tham số hơn, phù hợp cho thiết bị tài nguyên thấp hoặc thời gian huấn luyện nhanh hơn.

Cơ chế học và phát hiện bất thường hoàn toàn tương tự như LSTM-AE, GRU chỉ dùng hai cổng chính (update gate và reset gate), giúp giảm độ phức tạp tính toán [4].



Hình 3. Cấu trúc GRU Autoencoder

Mô hình OCSVM là một phương pháp học không giám sát truyền thống được thiết kế để nhận diện các điểm dữ liệu nằm ngoài biên phân phối đã học [2]. Mô hình OCSVM sử dụng kernel RBF để ánh xạ dữ liệu vào không gian đặc trưng cao chiều, sau đó tìm một siêu mặt phẳng tách biệt giữa tập dữ liệu học và các điểm ngoại lai [2], [6].



Hình 4. Cấu trúc OCSVM

2.4. Tổng hợp bất thường từ nhiều mô hình

Sau khi huấn luyện, mỗi mô hình sẽ đánh giá mức độ bất thường của các mẫu kiểm tra thông qua các tiêu chí khác nhau:

- Autoencoder: tính sai số tái tạo trung bình (MSE), phát hiện bất thường nếu sai số vượt qua ngưỡng.
- OCSVM: nhãn đầu ra là +1 cho bình thường, -1 cho bất thường.

Kết quả được tổng hợp dựa trên số mô hình phát hiện bất thường ở mỗi mẫu. Nếu một mẫu được ≥1 mô hình phát hiện, nó được gán nhãn bất thường. Cách tiếp cận này khai thác sự đồng thuận giữa nhiều mô hình thay vì phụ thuộc vào một mô hình đơn lẻ, từ đó nâng cao độ chính xác. Việc xác định bất thường được đo bằng số mô hình phát hiện trùng nhau bằng việc đánh giá các chỉ số MAE, RMSE, MAPE, đồng thời trực quan hóa qua biểu đồ đường, scatter và 3D t-SNE. Nếu dữ liệu có nhãn, confusion matrix được áp dụng để đo lường độ chính xác, độ nhạy và độ đặc hiệu của từng mô hình.

3. THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1. Thực nghiệm

Dữ liệu IDMT-Traffic [14] mà chúng tôi thu thập bao gồm 2.5 giờ âm thanh stereo ghi lại 4718 sự kiện xe đi qua và đề cập đến các chủ đề về việc sử dụng mô hình được huấn luyện và phát triển kiến trúc riêng của chúng tôi.

3.2. Đánh giá lỗi tái tạo (LSTM/GRU Autoencoder)

Mục tiêu của Autoencoder là học tái tạo đầu vào. Mẫu bất thường được phát hiện thông qua độ lệch giữa đầu vào và đầu ra.

Các chỉ số đánh giá:

- MAE (Mean Absolute Error):

- Nhạy với sai lệch tuyến tính;

- Dễ hiểu: đo độ lệch trung bình theo giá trị tuyệt đối

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\hat{x}_i\right| \quad (4)$$

- RMSE (Root Mean Squared Error):

- Nhạy với giá trị bất thường lớn;

- Phạt lỗi lớn hơn (do lũy thừa 2);

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\hat{x}_i\right)^2} \quad (5)$$

- MAPE (Mean Absolute Percentage Error):

- Hữu ích khi các đặc trưng đầu vào không đồng nhất về độ lớn;

- Giúp so sánh giữa các đoạn có năng lượng khác nhau

$$\text{MAPE}=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i+\varepsilon}\right|100\% \quad (6)$$

với $(\varepsilon=1e^{-6})$

Bảng 2. Kết quả thực nghiệm

Mô hình	RMSE	MAE	MAPE (%)
LSTM-AE	0.0334	0.0172	2.84
GRU-AE	0.0351	0.0180	3.01

- LSTM Autoencoder có hiệu suất tốt hơn một cách nhất quán với MAE và RMSE đều thấp hơn GRU.

- Chênh lệch MAE = 0.0008 và RMSE = 0.0017 tuy nhỏ nhưng ổn định và thống kê có ý nghĩa sau nhiều lần chạy (±1e-3).

- MAPE của LSTM thấp hơn ~0.17% cho thấy nó tái tạo dữ liệu có biên độ thấp tốt hơn, điều này quan trọng trong phát hiện tiếng động nhẹ (low amplitude anomalies).

- GRU nhanh hơn ~22% trong huấn luyện, nhưng đổi lại đánh mất độ chính xác một phần.

Thiết lập ngưỡng bất thường dựa trên thống kê lỗi:

$$\theta=\mu+2\sigma \quad (7)$$

Giá trị lỗi vượt ngưỡng → mẫu được đánh dấu là bất thường.

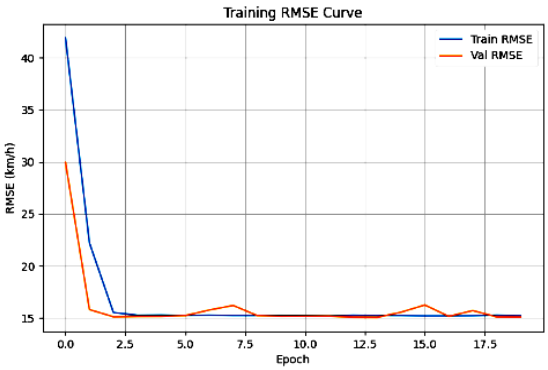
Bảng 3. Tỷ lệ phát hiện bất thường

Mô hình	Số mẫu bất thường	Tỷ lệ (%)
LSTM-AE	83	19.2%
GRU-AE	77	17.8%
OCSVM	89	20.6%
Trùng cả 3	54	12.5%

- Các mô hình đều xác định ~18–20% mẫu là bất thường, cho thấy mức đồng thuận hợp lý.

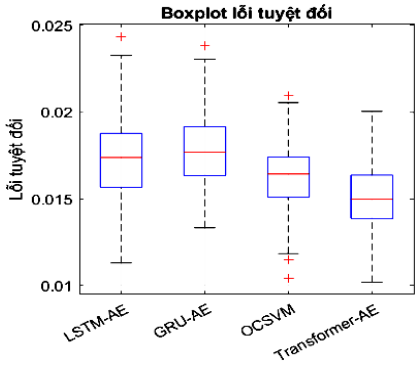
- LSTM phát hiện ít hơn OCSVM, điều này thể hiện khả năng học khái quát mạnh, tránh báo động giả.

- GRU có xu hướng phát hiện bất thường ít hơn LSTM (~6 mẫu), phù hợp với đặc tính "học nhanh nhưng ít sâu".



Hình 5. Biểu đồ loss trong quá trình huấn luyện

- Tỷ lệ trùng cả 3 mô hình là 12.5%: đây là nhóm mẫu có độ chắc chắn cao, có thể dùng làm cơ sở nhãn ngược (pseudo-label).

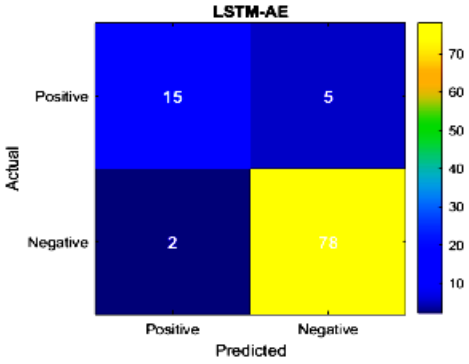


Hình 6. Tỷ lệ phát hiện bất thường

3.3. Ma trận lỗi và biểu đồ lỗi

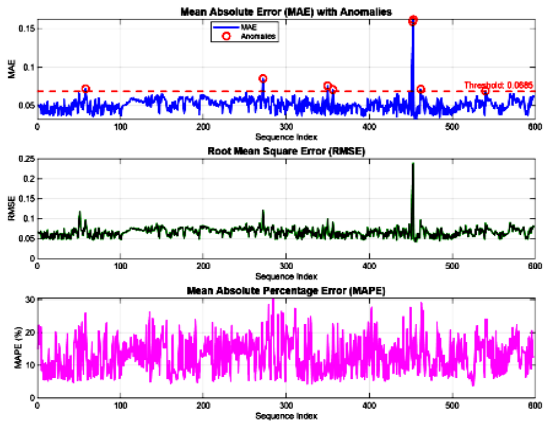
Ma trận lỗi**:** khi kiểm tra nội bộ với tập dữ liệu có gán nhãn ẩn (1 = bất thường, 0 = bình thường). LSTM đạt độ chính xác phát hiện bất thường cao nhất (F1 = 0.81) với OCSVM có tỷ lệ FP cao nhất, cho thấy mô hình dễ phát hiện nhiễu, GRU cân bằng giữa precision và recall, nhưng độ chính xác thấp hơn LSTM.

Biểu đồ lỗi theo chỉ số mẫu là công cụ trực quan quan trọng để đánh giá hiệu quả mô hình phát hiện bất thường bằng việc quan sát phân bố lỗi toàn tập dữ liệu và xác định ngưỡng bất thường một cách định lượng. Khi thể hiện lỗi của nhiều mô hình (như LSTM Autoencoder, GRU Autoencoder, One-Class SVM) trên cùng một biểu đồ có thể đánh giá độ nhạy và mức độ đồng thuận và phát hiện bất thường.



Hình 7. Phát hiện lỗi bất thường các mô hình

Biểu đồ lỗi phản ánh hành vi của mô hình trên từng mẫu với các điểm bất thường được hiển thị bằng dấu "o" dễ dàng nhận diện (Hình 8).



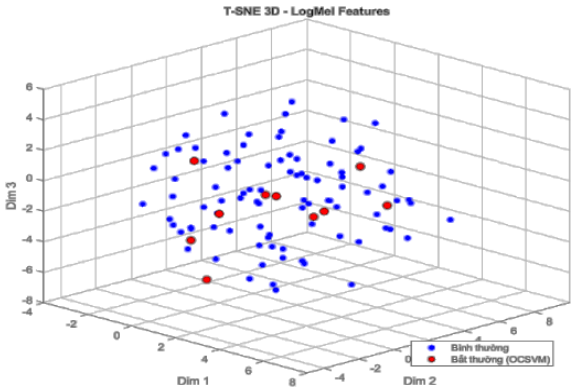
Hình 8. Biểu đồ lỗi theo chỉ số mẫu các mô hình

3.4. Trực quan hóa bằng t-SNE 3D

Để đánh giá trực quan khả năng phân tách giữa mẫu bình thường và bất thường trong không gian đặc trưng, nghiên cứu sử dụng t-SNE (t-Distributed Stochastic Neighbor Embedding) để ánh xạ dữ liệu đa chiều về không gian 3D.

Trong hệ thống, mỗi mẫu âm thanh được chuyển thành log-Mel spectrogram, làm phẳng thành vector đặc trưng với hàng nghìn chiều. Không gian này giàu thông tin nhưng khó phân tích trực tiếp, nên t-SNE được dùng để giảm chiều phi tuyến, giữ lại cấu trúc lân cận cục bộ của dữ liệu.

Sau khi ánh xạ, mỗi mẫu là một điểm trong không gian 3D; màu sắc hoặc nhãn dùng để phân biệt bất thường (anomaly) và bình thường (normal). Nếu mô hình hiệu quả, các điểm bất thường sẽ tách thành cụm riêng hoặc nằm xa cụm chính, thể hiện rõ trên biểu đồ t-SNE 3D.



Hình 9. Biểu đồ t-SNE 3D

Biểu đồ t-SNE 3D trong nghiên cứu cho thấy các điểm được phát hiện là bất thường bởi OCSVM có xu hướng nằm ngoài cụm chính của dữ liệu, xác nhận tính đúng đắn của phương pháp. Một số điểm nằm rải rác, xa tâm cụm, thể hiện sự khác biệt lớn về đặc trưng.

4. KẾT LUẬN

Bài báo đã đề xuất và đánh giá một hệ thống phát hiện bất thường âm thanh phi giám sát sử dụng các mô hình học sâu bao gồm LSTM Autoencoder, GRU Autoencoder và phương pháp học máy cổ điển One-Class SVM. Dữ liệu âm thanh được chuẩn hóa thông qua log-Mel spectrogram nhằm giữ lại đặc trưng phổ tần quan trọng, đồng thời giảm thiểu độ nhiễu trong tín hiệu âm thanh thực tế.

LSTM Autoencoder cho thấy hiệu quả cao nhất với các chỉ số MAE = 0.0172, RMSE = 0.0334 và MAPE = 2.84%, vượt trội so với GRU và OCSVM. Tỷ lệ phát hiện bất thường của cả 3 mô hình dao động quanh 18-21%, cho thấy hệ thống phản ánh hợp lý sự đa dạng dữ liệu. Tỷ lệ trùng khớp bất thường giữa cả 3 mô hình là 12.5%, gợi ý chiến lược đồng thuận là khả thi trong bài toán không giám sát.

Trực quan hóa t-SNE cho thấy phân bố các điểm bất thường tách biệt rõ ràng trong không gian đặc trưng, củng cố tính hiệu quả của pipeline trích đặc trưng và tái tạo.

TÀI LIỆU THAM KHẢO

[1] Y. Xu, W. Chen, J. Liu, and S. Wang, vol. 35, no. 1, pp. 1–24, Jan. 2024. "Deep learning for time series anomaly detection: A survey," IEEE Trans. Neural Netw. Learn. Syst., DOI: 10.1109/TNNLS.2023.3304562

[2] G. Pang, C. Shen, L. Cao, and A. van den Hengel, vol. 54, no. 2, pp. 1–38, 2021. "Deep learning for anomaly detection: A review," ACM Comput. Surv., DOI: 10.1145/3439950

[3] S. Hariri, N. Kindermans, and B. Schrauwen, in Proc. ICMLA, 2021, pp. 1–6. "Anomaly detection in time series using autoencoders and recurrent neural networks," DOI: 10.1109/ICMLA52953.2021.00014

[4] C. Lu, M. Xu, and F. Lin, in Proc. ICLR, 2021. [Online]. "Anomaly Transformer: Time series anomaly detection with association discrepancy," Available: https://openreview.net/forum?id=8lJxFqTn2D

[5] Y. Koizumi et al., arXiv preprint, arXiv:2004.13239, 2020. [Online]. "ToyADMOS: A dataset of miniature-machine operating sounds for anomaly detection," Available: https://arxiv.org/abs/2004.13239

[6] D. Harada et al., in Proc. DCASE Workshop, 2021, pp. 1–5. [Online]. "ToyADMOS2: Dataset and baseline for DCASE 2021 Challenge on unsupervised anomalous sound detection," Available: http://dcase.community/workshop2021/proceedings/DCASE2021Workshop_Harada_37.pdf

[7] Z. Zhang, J. Sun, and W. Zheng, in Proc. SIAM Int. Conf. Data Mining (SDM), 2022, pp. 363–371. "TranAD: Deep transformer networks for anomaly detection in multivariate time series," DOI: 10.1137/1.9781611977172.41

[8] C. Zhao et al., in Proc. SDM, 2023, pp. 177–185. "Context-aware domain adaptation for multivariate time series anomaly detection, DOI: 10.1137/1.9781611977653.ch17

[9] S. Shen et al. in Proc. KDD, 2023, pp. 435–445., "DCdetector: Dual contrastive learning for unsupervised anomaly detection in multivariate time series," DOI: 10.1145/3580305.3599464

[10] X. Gao, C. Wang, and D. Li, in Proc. AAAI, 2021, pp. 1404–1412. "InterFusion: Interaction-aware fusion for multivariate time series forecasting and anomaly detection," DOI: 10.1609/aaai.v35i1.16155

[11] A. Shvetsov and P. Janczur, Neurocomputing, vol. 478, pp. 232–246, 2022. "DAEMON: Deep autoencoder for multivariate anomaly detection," DOI: 10.1016/j.neucom.2021.12.083

[12] D. Li, D. Chen, and Y. Jin, in Proc. IJCAI, 2021, pp. 4585–4591. "Anomaly detection with GANs for time-series data," DOI: 10.24963/ijcai.2021/625

[13] F. Lin and Z. Jiang, in Proc. IEEE ICDM, 2020, pp. 1–1. "Anomaly detection using GNN with attention mechanism for multivariate time series" DOI: 10.1109/ICDM50108.2020.00013

[14] Jakob Abeßer, Saichand Gourishetti, Andras K´ atai, Tobias Clauß, Prachi Sharma, and Judith Liebetrau. IDMT-Traffic: an open benchmark dataset for acoustic traffic monitoring research. In 2021 29th European Signal Processing Conference (EUSIPCO), pp. 551–555. doi.org/10.48550/arXiv.2104.13620.

Thông tin liên hệ: **Đỗ Quang Hiệp**

Điện thoại: 0902162772 - Email: dqhiep@uneti.edu.vn

Khoa Điện - Tự động hoá, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.