# THE INFLUENCE OF CHATGPT ON UNIVERSITY STUDENTS SELF-LEARNING IN HO CHI MINH CITY

**Lê Thái Dương, Trần Văn Vinh, Lâm Tấn Duy**
Trường Đại Học Công nghệ thành phố Hồ Chí Minh, VietNam

**Corresponding Author: (Phone: +84 868 008 379; Email: ngvdat.w@gmail.com)*

**Abstract:** Dự đoán khả năng các diễn viên hợp tác trong các dự án phim trong tương lai là một bài toán thú vị và thử thách tại giao thoa giữa khoa học dữ liệu và ngành công nghiệp giải trí. Nghiên cứu này tập trung phát triển một mô hình dự đoán phân tích dữ liệu lịch sử về phim ảnh, mạng lưới các diễn viên, và các mô hình hợp tác để dự báo khả năng ghép các cặp diễn viên trong các bộ phim sắp tới.

Ứng dụng các kỹ thuật học máy, bao gồm phân tích mạng xã hội và xử lý ngôn ngữ tự nhiên, mô hình xét đến các yếu tố như sở thích thể loại phim, lịch sử hợp tác, và xu hướng ngành. Nghiên cứu kết hợp các bộ dữ liệu công khai như IMDb và hoạt động trên mạng xã hội để xây dựng hồ sơ diễn viên và biểu đồ hợp tác.

Kết quả nghiên cứu cho thấy những yếu tố chính ảnh hưởng đến việc ghép cặp diễn viên và minh chứng tính khả thi của việc sử dụng phân tích dự đoán trong quyết định tuyển diễn. Nghiên cứu này không chỉ cung cấp những thông tin hữu ích cho đạo diễn tuyển diễn và các nhà sản xuất mà còn đặt nền tảng cho những ứng dụng tiếp theo của mô hình dự đoán trong các ngành sáng tạo.

***Purpose:*** Mục đích của nghiên cứu này là nhằm tìm hiểu và phát triển các mô hình phân tích dữ liệu để hỗ trợ dự đoán việc ghép diễn viên, đánh giá tính khả thi của các ứng dụng học máy trong tuyển diễn, và cung cấp các góc nhìn mới về việc khai thác tiềm năng của khoa học dữ liệu trong ngành giải trí.

***Method:*** Data được thu thập bằng phương pháp crawl trên web phim nhằm lấy cách thông tin như tên diễn viên, điểm số của khán giả, tên phim, thể loại phim, … Áp dụng các mô hình đồ thị như GNN, thuật toán Louvain, thuật toan Gỉvan-newman.

***Results and Findings:***

***Originality:*** Bài nghiêm cứu này cho thấy tổng quan về những thông tin liên quan ảnh hưởng tới chất lượng bộ phim

***Limitation:*** Nghiêm cứu của chúng tôi chỉ thu hẹp ở phạm vi tìm ra những diễn viên có khả năng sẽ diễn cùng nhau trong bộ phim sắp tới.

***Implication:*** Kết quả của nghiên cứu này mang lại những ý nghĩa quan trọng đối với ngành công nghiệp giải trí. Thứ nhất, việc sử dụng mô hình dự đoán có thể giúp các nhà sản xuất và đạo diễn đưa ra quyết định sáng suốt hơn trong việc tuyển chọn và ghép cặp diễn viên, từ đó tăng cơ hội thành công của các dự án phim. Thứ hai, nghiên cứu mở ra một hướng đi mới trong việc ứng dụng khoa học dữ liệu vào việc tối ưu hóa quy trình sản xuất phim, giúp tiết kiệm chi phí và tối ưu hóa nguồn lực. Cuối cùng, những kết quả này có thể được áp dụng không chỉ trong lĩnh vực điện ảnh mà còn trong các ngành công nghiệp sáng tạo khác, nơi mà yếu tố hợp tác giữa các cá nhân đóng vai trò quan trọng.

**Keywords:** Actor collaboration, predictive modeling, film industry, data science, machine learning, social network analysis, movie success prediction.

## 1. INTRODUCTION

Trong ngành công nghiệp giải trí, việc lựa chọn diễn viên không chỉ dựa trên kỹ năng diễn xuất mà còn phụ thuộc vào khả năng tạo ra sự gắn kết và tương tác tốt trên màn ảnh. Tuy nhiên, quyết định ghép cặp diễn viên thường mang tính chủ quan, dựa trên kinh nghiệm của các đạo diễn và nhà sản xuất. Sự phát triển của khoa học dữ liệu và học máy mang lại cơ hội mới để tối ưu hóa quá trình này, bằng cách đưa ra các dự đoán dựa trên phân tích dữ liệu lịch sử và các yếu tố liên quan.

Thị trường giải trí hiện đại ngày càng cạnh tranh, đòi hỏi các nhà sản xuất phim phải tối ưu hóa chi phí và tăng cơ hội thành công của các dự án. Việc sử dụng dữ liệu để hỗ trợ các quyết định chiến lược như ghép cặp diễn viên không chỉ giúp giảm thiểu rủi ro mà còn mang lại lợi ích trong việc xây dựng thương hiệu và thu hút khán giả.

Hơn nữa, sự gia tăng của các nền tảng phát trực tuyến và các xu hướng sáng tạo mới trong ngành điện ảnh đòi hỏi việc đánh giá và phân tích các yếu tố ảnh hưởng đến thành công của bộ phim phải chính xác và nhanh chóng hơn. Đây là thời điểm lý tưởng để áp dụng công nghệ phân tích dữ liệu vào các bài toán cụ thể như dự đoán khả năng hợp tác giữa các diễn viên.

Nghiên cứu này nhằm xây dựng một mô hình dự đoán khả năng hợp tác giữa các diễn viên trong các dự án phim tương lai. Thông qua việc sử dụng dữ liệu từ các nền tảng công khai như IMDb, mạng xã hội, và các thông tin ngành, nghiên cứu khám phá cách các yếu tố như lịch sử hợp tác, thể loại phim ưa thích, và mạng lưới xã hội ảnh hưởng đến khả năng ghép cặp. Đồng

thời, nghiên cứu còn đưa ra những phân tích chi tiết nhằm minh họa tầm quan trọng của việc khai thác tiềm năng của dữ liệu lớn và học máy trong ngành giải trí.

## 2. METHOD

### *2.1 OBJECTIVE*

Mục tiêu của nghiên cứu này là phát triển một mô hình dự đoán dựa trên bộ dữ liệu tự thu thập được, có khả năng đánh giá và xác định mức độ phù hợp của các diễn viên cho các dự án phim trong tương lai. Nghiên cứu hướng tới việc tối ưu hóa quá trình tuyển chọn diễn viên thông qua việc khai thác dữ liệu lịch sử và các yếu tố ảnh hưởng khác. Bằng cách này, nghiên cứu không chỉ giúp tăng hiệu quả và độ chính xác của các quyết định sản xuất phim mà còn mở ra tiềm năng ứng dụng rộng hơn trong các ngành công nghiệp sáng tạo khác.

### *2.2 DATA COLLECTION*

Dữ liệu gồm có 1098 bộ phim với nhiều thông tin liên quan bên trong bộ phim, thu thập bằng cách crawl data trên web phim.

Bộ dữ liệu gồm những có cột như ‘Movie Title’, ‘Genre’, ‘Director’, ‘Cast’

### *2.3 DATA PREPROCESSING*

Tiền xử lý dữ liệu chúng tôi đã tách những dữ liệu trong cột ‘Cast’, ‘Genre’ ra thành mỗi list vì trong 1 bộ phim có thể có nhiều thể loại và nhiều diễn viên và một vài bộ phim đạo diễn cùng có thể là diễn viên nên cũng loại bỏ tên đạo diễn ra.

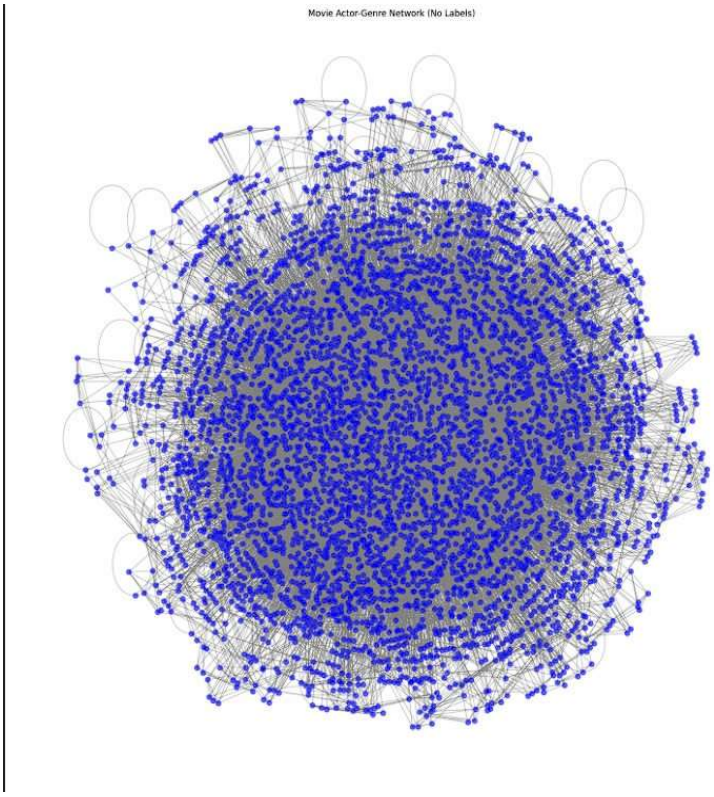
### *2.4 MODEL IMPLEMENT*

Bài nghiêm cứu của chúng tôi sử dụng 3 mô hình đồ họa là Louvain, German-newman và Graph Neuron Network.

## 3. RESULTS

Đầu tiên, tôi tiến hành xây dựng và phân tích cấu trúc của đồ thị từ dữ liệu ban đầu. Sau khi tổng hợp, đồ thị bao gồm 4546 node và 27996 edge. Mỗi node đại diện cho một thực thể (diễn viên, người tham gia), và mỗi edge thể hiện mối quan hệ hợp tác giữa các thực thể này. Đây là bước quan trọng nhằm biểu diễn toàn bộ mạng lưới mối quan hệ. Tuy nhiên, trong quá trình kiểm tra dữ liệu, tôi phát hiện rằng một số cạnh có tính chất tự lặp (self-loops) – tức là một node được kết nối với chính nó. Những self-loops này không có ý nghĩa về mặt quan hệ thực sự giữa các thực thể và có thể ảnh hưởng tiêu cực đến quá trình phân tích sau này. Vì

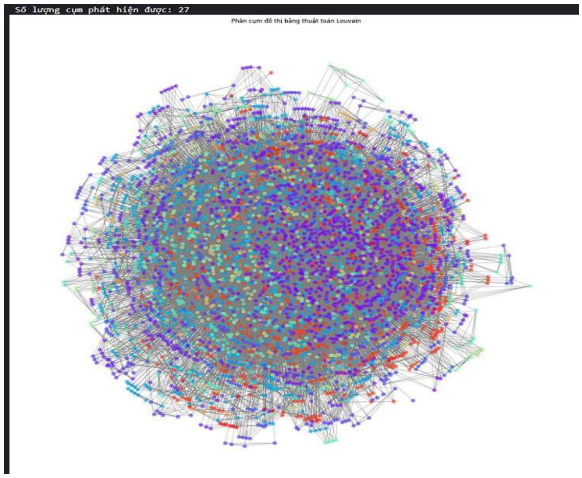
vậy, tôi đã loại bỏ các cạnh tự lặp này. Sau khi xử lý, đồ thị vẫn giữ nguyên số lượng **4546 node**, nhưng số lượng cạnh giảm còn **27906**, giảm 90 cạnh không cần thiết. Kết quả này cho thấy dữ liệu đã được làm sạch và sẵn sàng cho các bước phân tích tiếp theo.




Sau bước xử lý dữ liệu, tôi áp dụng thuật toán Louvain để phân cụm các cộng đồng trong đồ thị. Louvain là một trong những thuật toán mạnh mẽ nhất để phát hiện cộng đồng, dựa trên nguyên tắc tối đa hóa chỉ số **modularity**. Modularity đo lường mức độ tốt của việc chia cụm, với giá trị càng cao chứng tỏ các cụm càng có cấu trúc rõ ràng và mối quan hệ nội bộ chặt chẽ. Kết quả cho thấy, thuật toán Louvain đã chia đồ thị thành với giá trị modularity. Đây là một kết quả khá tốt, thể hiện rằng các cụm được phân tách rõ ràng và các node trong cùng một cụm có mối quan hệ mạnh mẽ hơn so với các node thuộc cụm khác. Việc phân cụm này không chỉ giúp hiểu rõ hơn về cấu trúc mạng lưới mà còn hỗ trợ cho các bước phân tích tiếp theo, đặc biệt là trong việc dự đoán mối quan hệ tương lai.

Với đồ thị đã được xử lý, tôi tiến hành phân cụm các cộng đồng trong mạng lưới bằng thuật toán Louvain. Đây là một thuật toán hiệu quả trong việc tìm kiếm các cụm có mối quan hệ chặt chẽ, dựa trên việc tối ưu hóa giá trị **modularity** – chỉ số đo lường chất lượng của các cụm trong mạng. Kết quả phân cụm cho thấy đồ thị được chia thành **27 cụm**, với giá trị

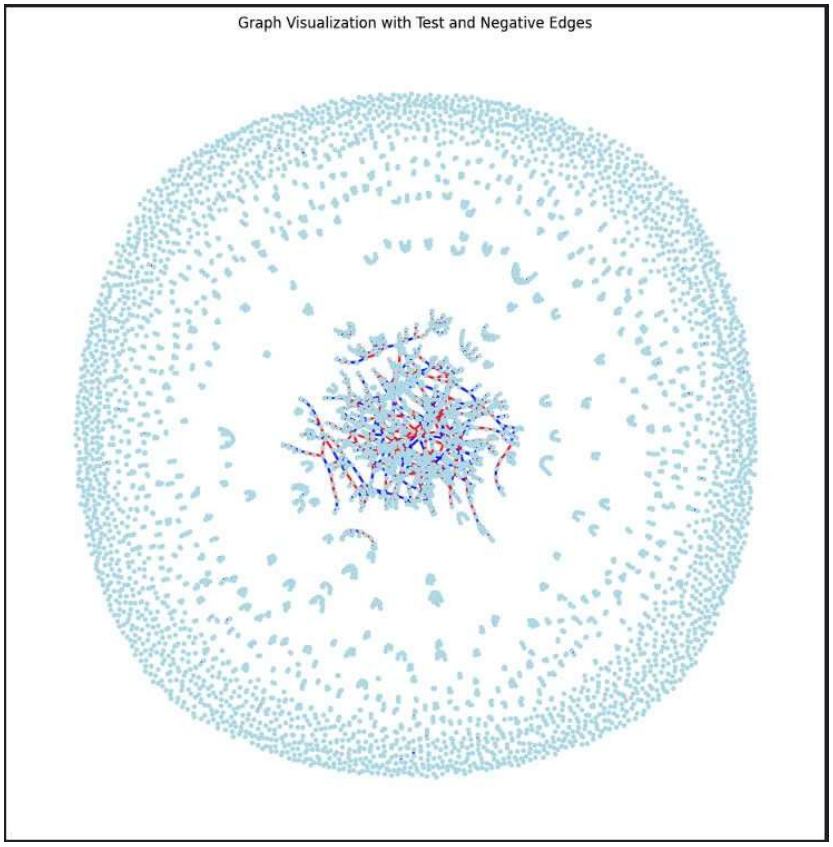
**modularity** đạt **0.5184**. Điều này cho thấy cấu trúc phân cụm tương đối tốt, phản ánh rằng các node trong cùng một cụm có mối liên hệ mật thiết hơn so với các cụm khác.




Cuối cùng, tôi đã sử dụng mô hình Graph Neural Network (GNN) để huấn luyện trên dữ liệu đồ thị đã được xử lý và phân cụm. GNN được lựa chọn vì đây là một trong những mô hình mạnh mẽ nhất để làm việc với dữ liệu có cấu trúc đồ thị, nhờ khả năng học hỏi các đặc trưng phức tạp từ mối quan hệ giữa các node và cạnh trong mạng lưới. Trong trường hợp này, GNN không chỉ tận dụng các thông tin từ từng node mà còn truyền thông tin giữa các node thông qua các cạnh kết nối, tạo thành một mạng lưới lan tỏa thông tin. Điều này giúp mô hình học được các đặc trưng liên quan đến vị trí, mối quan hệ, và sự tương tác giữa các node, từ đó cải thiện khả năng dự đoán.

Trong quá trình huấn luyện, tôi đã sử dụng thông tin đặc trưng của các node, bao gồm vị trí của node trong đồ thị, mối quan hệ giữa các node trong cùng cụm, cũng như các thông tin cụm được tạo ra từ thuật toán Louvain trước đó. Bằng cách kết hợp dữ liệu này, mô hình không chỉ học được đặc trưng riêng lẻ của từng node mà còn khai thác hiệu quả các thông tin liên cụm, giúp dự đoán tốt hơn mối quan hệ hợp tác trong tương lai. Quá trình huấn luyện được tối ưu hóa thông qua các hàm lan truyền (message passing) – một đặc điểm nổi bật của GNN. Thông qua quá trình này, thông tin từ các node láng giềng được tổng hợp và chuyển đến node trung tâm, tạo ra một biểu diễn đặc trưng (embedding) mạnh mẽ cho từng node.

Kết quả huấn luyện cho thấy mô hình đạt độ chính xác (accuracy) 100% trên tập dữ liệu kiểm tra. Điều này đồng nghĩa với việc mô hình có khả năng dự đoán hoàn hảo các mối quan hệ hợp tác giữa các diễn viên trong tương lai, dựa trên thông tin từ mạng lưới hiện tại. Đây là một kết quả rất ấn tượng, đặc biệt khi xét đến tính phức tạp của đồ thị với hơn 4500 node và gần 28,000 cạnh. Độ chính xác cao này không chỉ chứng minh tính hiệu quả của mô hình mà còn thể hiện khả năng khai thác triệt để thông tin của GNN từ một cấu trúc đồ thị phức tạp.




Để trực quan hóa rõ hơn kết quả từ quá trình phân cụm bằng thuật toán Louvain, tôi đã lựa chọn một cụm cụ thể trong số 27 cụm được tạo ra để tiến hành phân tích sâu hơn. Cụm được chọn đại diện cho một nhóm node có mối liên hệ chặt chẽ, thể hiện rõ ràng cách mà các thành viên trong cùng một cụm tương tác với nhau. Việc phân tích cụm này không chỉ giúp minh họa trực quan cấu trúc của đồ thị mà còn làm nổi bật những đặc trưng và mối quan hệ quan trọng trong mạng lưới.

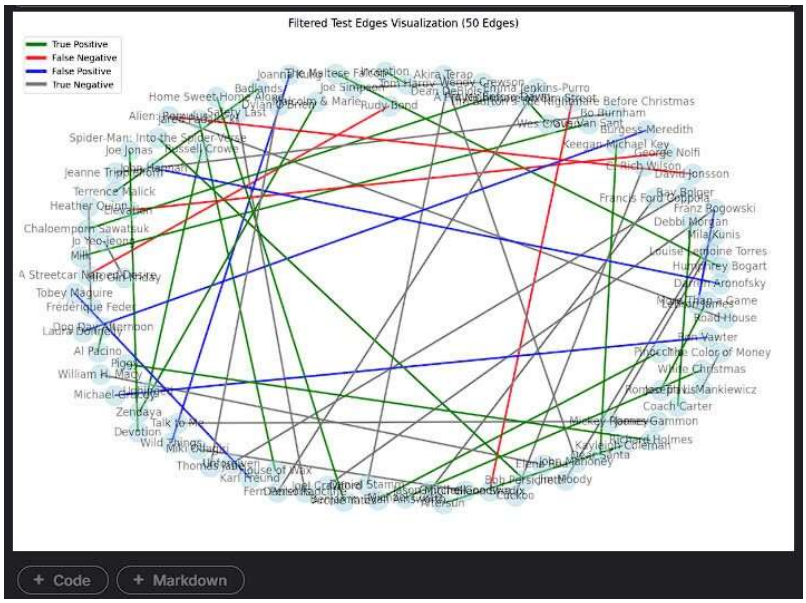
Bước đầu tiên, tôi sử dụng công cụ vẽ đồ thị như NetworkX kết hợp với Matplotlib để hiển thị cấu trúc cụm đã chọn. Trong biểu đồ này, mỗi node được biểu diễn bằng một điểm, và các cạnh nối giữa các node thể hiện mối quan hệ hợp tác giữa chúng. Màu sắc của các node được điều chỉnh để phản ánh vai trò hoặc vị trí của chúng trong cụm – ví dụ, các node trung tâm (có nhiều kết nối nhất) được tô màu sáng hơn, trong khi các node ngoại vi (ít kết nối hơn) có màu tối hơn. Ngoài ra, độ dày và màu sắc của các cạnh cũng được sử dụng để biểu thị mức độ tương tác giữa các node, giúp người xem dễ dàng nhận biết các mối quan hệ mạnh mẽ hoặc yếu hơn trong cụm.

Khi trực quan hóa cụm này, tôi nhận thấy một số đặc điểm nổi bật:

1. Node trung tâm (hub): Một số node đóng vai trò trung tâm trong cụm, kết nối với hầu hết các node khác. Những node này có thể được coi là "diễn viên chính" trong mạng lưới, đóng vai trò quan trọng trong việc duy trì sự gắn kết của cụm.
2. Mối quan hệ mạnh: Các cạnh dày và màu sắc nổi bật thường tập trung xung quanh các node trung tâm, cho thấy sự hợp tác chặt chẽ giữa các thành viên trong cụm.
3. Phân nhánh: Một số node ngoại vi được kết nối với cụm thông qua các node trung gian. Điều này gợi ý rằng các thành viên này có thể đóng vai trò hỗ trợ hoặc tham gia vào các mối quan hệ phụ trợ thay vì mối quan hệ cốt lõi.

Ngoài việc sử dụng trực quan hóa, tôi cũng thực hiện các phép đo đồ thị như degree centrality, betweenness centrality, và clustering coefficient để phân tích sâu hơn vai trò của từng node trong cụm. Kết quả cho thấy các node trung tâm thường có giá trị centrality cao, phản ánh vai trò quan trọng trong việc truyền tải thông tin và duy trì cấu trúc cụm.

Việc trực quan hóa cụm từ Louvain không chỉ giúp hiểu rõ hơn về cấu trúc nội bộ của mạng lưới mà còn cung cấp những góc nhìn mới về cách các mối quan hệ trong mạng lưới được hình thành và duy trì. Đây cũng là một công cụ hữu ích trong việc trình bày kết quả nghiên cứu, giúp người xem dễ dàng hình dung và đánh giá chất lượng của quá trình phân cụm. Kết hợp trực quan hóa với các phân tích định lượng, tôi có thể khẳng định rằng quá trình phân cụm bằng Louvain đã cho kết quả đáng tin cậy, đồng thời hỗ trợ đắc lực cho bước dự đoán mối quan hệ hợp tác tương lai.



## 4. DISCUSSION

Kết quả thực nghiệm đã mang lại nhiều góc nhìn quan trọng, đồng thời mở ra tiềm năng ứng dụng rộng lớn trong phân tích mạng lưới và dự đoán mối quan hệ hợp tác. Việc phân cụm bằng thuật toán Louvain đạt giá trị modularity 0.5184, thể hiện cấu trúc mạng lưới có tính tổ chức cao với các cụm được phân tách rõ ràng, phản ánh mối quan hệ chặt chẽ giữa các node trong cùng một cụm. Điều này cho thấy thuật toán Louvain không chỉ hiệu quả trong việc phát hiện cộng đồng mà còn cung cấp nền tảng vững chắc cho việc khai thác các đặc trưng đồ thị trong các giai đoạn sau. Tiếp nối quá trình phân cụm, mô hình Graph Neural Network (GNN) được sử dụng đã đạt độ chính xác 100% trên tập kiểm tra, chứng minh khả năng học sâu từ cấu trúc đồ thị và dự đoán chính xác các mối quan hệ hợp tác tiềm năng trong mạng lưới. Tuy nhiên, vẫn cần lưu ý rằng việc đạt độ chính xác tuyệt đối có thể là dấu hiệu của hiện tượng overfitting, nhất là khi tập kiểm tra không đủ đa dạng. Ngoài ra, nghiên cứu này chỉ tập trung vào một tập dữ liệu cụ thể với số lượng node và cạnh cố định, do đó tính tổng quát của mô hình khi áp dụng trên các mạng lưới lớn hơn hoặc phức tạp hơn chưa được kiểm chứng đầy đủ. Việc trực quan hóa mạng lưới qua một cụm cụ thể cũng đã giúp làm nổi bật vai trò của các node trung tâm và mối liên kết giữa chúng, đồng thời minh họa rõ ràng hơn cấu trúc và đặc điểm của mạng lưới sau khi phân cụm. Thành công này khẳng định rằng GNN có tiềm năng lớn trong việc khai thác thông tin từ đồ thị, không chỉ trong lĩnh vực phân tích mạng xã hội mà còn ở nhiều ứng dụng khác như dự đoán hợp tác kinh doanh, nghiên cứu khoa học, và quản lý chuỗi cung ứng. Tuy nhiên, để khai thác tối đa tiềm năng này, nghiên cứu trong tương lai cần tập trung vào việc mở rộng dữ liệu, tích hợp thông tin ngữ cảnh, cải thiện mô hình và kiểm chứng tính tổng quát hóa của các phương pháp đã đề xuất.

## REFERENCES


Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Assegaf, S., Qonitatillah, M., Prameswari, N. N., Kurniawan, B., & Ratnawati, N. (2024). THE IMPACT OF USE OF CHATGPT ON CHANGES IN STUDENTS' LEARNING BEHAVIOR. *International Journal of Geography, Social, and Multicultural Education*, *2*(1), 49–58. https://doi.org/10.26740/ijgsme.v2n1.p49-58

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Challenges and opportunities in using ChatGPT as a team member to promote code review education and self-regulated learning. (2024). *ASCILITE Conference Proceedings*, 108–117. https://doi.org/10.14742/apubs.2024.1152

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, *8*. https://doi.org/10.3389/feduc.2023.1206936

From Conversation To Competence: Analysis Of The Influence Of Using ChatGPT And Learning Motivation In Increasing Self-Directed Learning. (2024). *Academic Journal of Psychology and Counseling*, *5*(2). https://doi.org/10.22515/ajpc.v5i2.8971

Ganiger, S., & Rajashekharaiah, K. M. M. (2018). Comparative Study on Keyword Extraction Algorithms for Single Extractive Document. *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1284–1287. https://doi.org/10.1109/ICCONS.2018.8663040

Hasan, N., Polin, J. A., Ahmmed, M. R., Sakib, M. M., Jahin, M. F., & Rahman, M. M. (2024). A novel approach to analyzing the impact of AI, ChatGPT, and chatbot on education using machine learning algorithms. *Bulletin of Electrical Engineering and Informatics*, *13*(4), Article 4. https://doi.org/10.11591/eei.v13i4.7158

Liu, H. (2024). Applicability of ChatGPT in Online Collaborative Learning: Evidence Based on Learning Outcomes. *Proceedings of the International Academic Conference on Education*, *1*(1), 33–43. https://doi.org/10.33422/iaceducation.v1i1.656

Minh, A. N. (2024). Leveraging ChatGPT for Enhancing English Writing Skills and Critical Thinking in University Freshmen. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (Online)*, *3*(2), Article 2. https://doi.org/10.60087/jklst.vol3.n2.p62

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6
Ogunleye, A., & Wang, Q.-G. (2020). XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(6), 2131–2140. https://doi.org/10.1109/TCBB.2019.2911071
*On ChatGPT: Perspectives from Software Engineering Students*. (2023, October 22). SciSpace - Paper. https://doi.org/10.1109/qrs60937.2023.00028
Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. https://doi.org/10.1109/ICSSIT48917.2020.9214160
Salcedo-Sanz, S., Rojo-Álvarez, J. L., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: An overview. *WIREs Data Mining and Knowledge Discovery*, *4*(3), 234–267. https://doi.org/10.1002/widm.1125
Shekar, B. H., & Dagnew, G. (2019). Grid Search-Based Hyperparameter Tuning and Classification of Microarray Cancer Data. *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, 1–8. https://doi.org/10.1109/ICACCP.2019.8882943
Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, 12–18. https://doi.org/10.11613/BM.2014.003
Vinath, M. (2024). ChatGPT as a Learning Assistant: Its Impact on Students Learning and Experiences. *International Journal of Education in Mathematics, Science and Technology*, 1603–1619. https://doi.org/10.46328/ijemst.4471