


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/7BQmKw8UQfE>
- Link slides (dạng .pdf đặt trên Github):  
<https://github.com/nguyendt-tn/CS2205.APR2024>
- Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới
- Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in

<ul style="list-style-type: none"><li>• Họ và Tên: Dương Thanh Nguyên</li><li>• MSSV: 230201048</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS2205.APR2024</li><li>• Tự đánh giá (điểm tổng kết môn): 7/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân:</li><li>• Link Github: <a href="https://github.com/nguyendt-tn/CS2205.APR2024">https://github.com/nguyendt-tn/CS2205.APR2024</a></li></ul>
--	---

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

NHẬN DIỆN CẢM XÚC QUA GIỌNG NÓI SỬ DỤNG  
KOLMOGOROV–ARNOLD NETWORKS

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

SPEECH EMOTION RECOGNITION USING KOLMOGOROV–ARNOLD  
NETWORK

## TÓM TẮT *(Tối đa 400 từ)*

Trong bối cảnh công nghệ thông tin ngày càng phát triển, việc nhận diện cảm xúc qua giọng nói trở thành một lĩnh vực nghiên cứu có tiềm năng lớn, đặc biệt là trong việc cải thiện tương tác với người dùng và hỗ trợ tư vấn sức khỏe tâm lý từ xa.

Nghiên cứu khởi đầu bằng việc đánh giá các giới hạn của mô hình nhận diện cảm xúc hiện hành và đề xuất một giải pháp sử dụng Kolmogorov–Arnold Networks (KANs) để tăng độ chính xác và hiệu suất. KANs là một kiến trúc mới trong học sâu, hứa hẹn mang lại những cải tiến đáng kể bởi khả năng tính toán về toán học vượt trội, chính xác hơn.

Nghiên cứu này tập trung vào việc phát triển, tinh chỉnh và thử nghiệm mô hình Mạng Kolmogorov–Arnold (KANs) trong một bối cảnh thực tế. Không chỉ nhằm đến việc tăng cường độ chính xác trong quá trình phân tích cảm xúc qua giọng nói mà còn khai thác tiềm năng ứng dụng rộng rãi của công nghệ này trong các hoàn cảnh khác nhau.

Ngoài ra, nghiên cứu này còn hướng đến việc xây dựng một nền tảng vững chắc cho các ứng dụng trong tương lai, nơi mà sự tương tác giữa con người và máy móc trở nên tự nhiên và nhạy bén hơn. Sự kết hợp giữa mô hình KANs và các phương pháp xử lý giọng nói không chỉ mở ra nhiều cơ hội mới trong lĩnh vực giao tiếp thông minh, mà còn góp phần tạo nên những giải pháp hỗ trợ tâm lý và chăm sóc sức khỏe tinh thần từ xa, đáp ứng nhu cầu ngày càng cao của xã hội hiện đại.

## **GIỚI THIỆU** (Tối đa 1 trang A4)

Nhận diện cảm xúc qua giọng nói là một công nghệ hứa hẹn giúp máy tính hiểu và tương tác với con người hiệu quả và tự nhiên hơn. Công nghệ này đặc biệt hữu ích trong các lĩnh vực như tổng đài chăm sóc khách hàng, theo dõi sức khỏe tâm lý, và các thiết bị thông minh trong nhà và văn phòng. Trong bối cảnh gia tăng các vấn đề tâm lý, đặc biệt là sau đại dịch COVID-19, nhu cầu về các giải pháp hỗ trợ tâm lý đã tăng cao do tình trạng thiếu hụt chuyên gia. Chính vì vậy, các giải pháp nhận diện cảm xúc qua giọng nói trở nên ngày càng cần thiết và có giá trị. Nghiên cứu này tập trung vào việc áp dụng và cải thiện các kỹ thuật học máy và học sâu để phát triển một mô hình nhận diện cảm xúc hiệu quả qua giọng nói.

Các nghiên cứu trước đây dùng kỹ thuật học máy truyền thống để trích xuất đặc trưng âm thanh, nhưng chưa cho thấy hiệu quả. Các kỹ thuật học sâu hiện đã thể hiện hiệu quả vượt trội hơn trong nhiều nhiệm vụ xử lý hình ảnh và ngôn ngữ tự nhiên.

Trong nghiên cứu này chủ yếu sử dụng học sâu để nhận diện cảm xúc qua giọng nói. Chúng tôi cũng giới thiệu Kolmogorov–Arnold Networks (KANs)[1], một mô hình mạng nơ-ron tiên tiến, lấy cảm hứng từ định lý biểu diễn Kolmogorov-Arnold, mang lại độ chính xác và hiệu quả cao hơn mô hình MLP truyền thống, cải thiện hiệu suất nhận diện

Input:

- Tín hiệu âm thanh: Tín hiệu âm thanh ghi lại giọng nói của người nói.
- Dữ liệu ngữ cảnh: Dữ liệu ngữ cảnh có thể bao gồm văn bản lời nói, hình ảnh video, hoặc thông tin về người nói và chủ đề giao tiếp.

Output:

- Cảm xúc: Cảm xúc của người nói được xác định dưới dạng một hoặc nhiều nhãn cảm xúc như vui, buồn, tức giận, sợ hãi, v.v.
- Mức độ tin cậy: Mức độ tin cậy của kết quả nhận diện cảm xúc. Nhận cảm xúc từ giọng nói.



## MỤC TIÊU

*(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)*

- Cải thiện độ chính xác trong nhận diện cảm xúc qua giọng nói sử dụng KANs
- Khám phá hiệu quả của KANs trong ứng dụng thực tế
- Xây dựng mô hình tối ưu cho nhận diện cảm xúc qua giọng nói từ KANs

## NỘI DUNG VÀ PHƯƠNG PHÁP

*(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*

### Nội Dung:

- Nghiên cứu và phân tích đặc trưng giọng nói để phân biệt cảm xúc.
- Xây dựng và so sánh hiệu suất của mô hình KANs với các mô hình truyền thống (LSTM, CNN, RNN).
- Đánh giá khả năng và hiệu quả của KANs trong các tình huống thực tế và môi trường âm thanh đa dạng.
- Tối ưu hóa kiến trúc và các tham số của mô hình KANs để đạt được độ chính xác tốt nhất trong nhận diện cảm xúc qua giọng nói.

### Phương Pháp Thực Hiện:

- Sử dụng các tập dữ liệu giọng nói đã được gán nhãn cảm xúc như RAVDESS, IEMOCAP
- Tiền xử lý dữ liệu bao gồm loại bỏ nhiễu và trích xuất các đặc trưng như MFCCs, formants, pitch.
- Xây dựng mô hình sử dụng KANs và huấn luyện trên dữ liệu đã tiền xử lý.

- Đánh giá hiệu suất bằng các chỉ số như accuracy, precision, recall, F1-score và so sánh với các mô hình truyền thống (LSTM, CNN, RNN).
- Xác định các tình huống ứng dụng thực tế và thực hiện thử nghiệm trên các dữ liệu giọng nói từ các môi trường âm thanh khác nhau.
- Tối ưu hóa kiến trúc và tham số của mô hình sử dụng KANs bằng cách thử nghiệm và điều chỉnh các yếu tố như số lớp, số nơ-ron, hàm kích hoạt.
- Sử dụng các kỹ thuật huấn luyện tiên tiến như learning rate annealing, early stopping, và k-fold cross-validation và giảm thiểu tài nguyên tính toán để tối ưu hiệu suất và thời gian huấn luyện của mô hình.

## KẾT QUẢ MONG ĐỢI

*(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*

- KANs mang lại độ chính xác cao hơn và hiệu suất tốt hơn so với các phương pháp truyền thống (LSTM, CNN, RNN) trong nhận diện cảm xúc qua giọng nói, độ chính xác hơn 85% trên bộ dữ liệu RAVDESS và hơn 75% trên bộ dữ liệu IEMOCAP.
- KANs có khả năng ứng dụng hiệu quả trong các tình huống thực tế, duy trì hiệu suất tốt trong các môi trường âm thanh phức tạp.
- Xây dựng được mô hình KANs tối ưu với hiệu suất tính toán cao và tốc độ huấn luyện nhanh, dễ dàng triển khai và ứng dụng.

## TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- [1].Ziming Liu and Yixuan Wang and Sachin Vaidya and Fabian Ruehle and James Halverson and Marin Soljačić and Thomas Y. Hou and Max Tegmark (2024). KANs: Kolmogorov-Arnold Networks. arXiv preprint arXiv:2404.19756.
- [2].Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.
- [3]. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.

- [4]. Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J.M., Fernández-Martínez, F. (2021). A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset. Appl. Sci. 2022, 12, 327, doi: 10.3390/app12010327.
- [5]. Boigne, J., Liyanage, B., Östrem, T. (2020). Recognizing more emotions with less data using self-supervised transfer learning. arXiv preprint arXiv:2011.05585.
- [6]. Pepino, L., Riera, P., Ferrer, L. (2021). Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. Proc. Interspeech 2021, 3400-3404, doi: 10.21437/Interspeech.2021-703.
- [7]. Chen, L. W., Rudnicky, A. (2021). Exploring Wav2vec 2.0 fine-tuning for improved speech emotion recognition. arXiv preprint arXiv:2110.06309.