

TRƯỜNG ĐẠI HỌC MỎ ĐỊA CHẤT

KHOA CÔNG NGHỆ THÔNG TIN

-----o0o-----

BÁO CÁO SINH HOẠT HỌC THUẬT

**Đề tài: Nghiên cứu mô hình Foundation Models và Temporal
Fusion Transformer trong dự báo chuỗi thời gian đa biến**

Người báo cáo : Nguyễn Thu Hằng

Đơn vị : Bộ môn Tin học Kinh tế

Hà Nội, 06/ 2026

MỤC LỤC

1	GIỚI THIỆU	3
1.1	Bối cảnh nghiên cứu.....	3
1.2	Mục tiêu và phạm vi nghiên cứu	4
1.2.1	Mục tiêu nghiên cứu.....	4
1.2.2	Phạm vi nghiên cứu.....	5
2	CƠ SỞ LÝ THUYẾT.....	6
2.1	Bài toán dự báo chuỗi thời gian đa biến.....	6
2.2	Kiến trúc Transformer và ứng dụng cho chuỗi thời gian	6
2.3	Temporal Fusion Transformer (TFT).....	8
2.4	Foundation Models cho chuỗi thời gian.....	10
2.4.1	Tổng quan.....	10
2.4.2	TimesFM	10
2.4.3	Chronos-2	11
2.4.4	MOIRAI	12
2.5	Kỹ thuật học chuyển tiếp (Transfer Learning)	13
2.5.1	Tổng quan.....	13
2.5.2	Các chiến lược fine-tuning	14
3	THỰC NGHIỆM VÀ ĐÁNH GIÁ	14
3.1	Tổng quan thiết kế thực nghiệm.....	14
3.2	Case Study 1 – Dự báo phụ tải điện với TFT Transfer Learning	
	15	

3.3	Case Study 2 – Dự báo giá cổ phiếu VN30 với Foundation Models	19
4	KẾT LUẬN.....	20
5	TÀI LIỆU THAM KHẢO.....	22

1 GIỚI THIỆU

1.1 Bối cảnh nghiên cứu

Dự báo chuỗi thời gian đa biến (multivariate time series forecasting – MTSF) là một bài toán trung tâm trong khoa học dữ liệu ứng dụng, với phạm vi ảnh hưởng rộng lớn từ quản lý năng lượng, tài chính đến y tế và vận tải. Đặc trưng nổi bật và đồng thời là thách thức chính của dự báo chuỗi thời gian đa biến nằm ở sự tương tác phi tuyến giữa nhiều biến theo thời gian: mô hình phải đồng thời nắm bắt các mẫu phụ thuộc dài hạn và ngắn hạn trong cùng một chuỗi, mối quan hệ tương quan chéo giữa các biến khác nhau, và ảnh hưởng định lượng của các biến ngoại sinh đã biết trước như lịch, thời tiết, sự kiện vĩ mô [1].

Lĩnh vực này đã trải qua ba làn sóng phát triển lớn. Làn sóng đầu tiên thuộc về mô hình thống kê kinh điển: ARIMA, VAR, Holt-Winters – có nền tảng lý thuyết vững chắc nhưng bị giới hạn bởi giả định tuyến tính [2]. Làn sóng thứ hai được dẫn dắt bởi học sâu: LSTM, GRU, Seq2Seq với attention – phá vỡ giới hạn tuyến tính nhưng còn hạn chế trong xử lý phụ thuộc cực dài hạn và yêu cầu lượng dữ liệu lớn [3]. Làn sóng thứ ba – đang diễn ra – được định hình bởi kiến trúc Transformer và Foundation Models: các mô hình quy mô lớn được huấn luyện trên dữ liệu đa dạng khổng lồ, có khả năng dự báo zero-shot hoặc few-shot trên miền dữ liệu mới [4].

Trong bối cảnh Việt Nam, nhu cầu ứng dụng MTSF ngày càng cấp thiết. Về phía ngành điện, Tập đoàn Điện lực Việt Nam (EVN) phải lập kế hoạch vận hành hệ thống điện với tổng công suất lắp đặt vượt 80 GW trong bối cảnh nhu cầu điện tăng trưởng bình quân 8–10%/năm và tỷ lệ năng lượng tái tạo tăng mạnh – tạo ra tính không chắc chắn lớn hơn trong dự báo phụ tải so với hệ thống điện truyền thống [5]. Về phía tài chính, thị trường chứng khoán Việt Nam đã đạt hơn 7 triệu tài khoản nhà đầu tư cá nhân và tổng vốn hóa gần 200 tỷ USD vào cuối năm 2024, đặt ra yêu cầu cấp thiết về công cụ phân tích và dự báo giá độ chính xác cao.

1.2 Mục tiêu và phạm vi nghiên cứu

1.2.1 Mục tiêu nghiên cứu

Nghiên cứu hướng đến mục tiêu tổng quát là đánh giá có hệ thống và so sánh toàn diện hai hướng tiếp cận tiên tiến trong dự báo chuỗi thời gian đa biến – mô hình Temporal Fusion Transformer kết hợp học chuyển tiếp và các Foundation Models chuyên biệt cho chuỗi thời gian – nhằm xác định điều kiện, phạm vi và mức độ hiệu quả của từng hướng trong bối cảnh dữ liệu thực tiễn tại Việt Nam. Từ đó, nghiên cứu xây dựng cơ sở lý luận và thực nghiệm để đề xuất khung phương pháp luận hỗ trợ việc lựa chọn mô hình phù hợp cho các bài toán MTSF trong điều kiện đặc thù của các thị trường mới nổi – nơi dữ liệu thường có quy mô hạn chế, tính không đồng nhất cao và các mẫu thời gian mang đặc điểm địa phương rõ rệt.

Để đạt được mục tiêu tổng quát trên, nghiên cứu triển khai bốn mục tiêu cụ thể:

(1) Phân tích chuyên sâu kiến trúc, cơ chế hoạt động và đặc điểm kỹ thuật của các mô hình nghiên cứu: TFT, TimeGPT-1, Chronos, Lag-Llama và MOIRAI, làm cơ sở lý thuyết cho toàn bộ nghiên cứu.

(2) Đánh giá hiệu quả của kỹ thuật học chuyển tiếp (transfer learning) áp dụng cho TFT, đặc biệt tập trung vào chiến lược partial fine-tuning, ablation study các phương án đóng băng tham số và phân tích ngưỡng tương đồng miền (domain similarity) cần thiết để TL mang lại lợi ích thực sự.

(3) Đánh giá và so sánh khả năng tổng quát hóa zero-shot và hiệu quả fine-tuned của các Foundation Models (TimeGPT-1, Chronos) trên dữ liệu ngoài phân phối huấn luyện, phân tích mức độ cải thiện của fine-tuning so với zero-shot theo đặc điểm chuỗi thời gian.

(4) Xây dựng khung so sánh và hướng dẫn thực hành về lựa chọn mô hình dựa trên các tiêu chí: quy mô dữ liệu, yêu cầu covariates, yêu cầu giải thích được, ràng buộc tính toán và bảo mật dữ liệu.

1.2.2 Phạm vi nghiên cứu

Về đối tượng nghiên cứu, bài seminar tập trung vào bài toán dự báo chuỗi thời gian đa biến có giám sát (supervised multivariate time series forecasting) theo cơ chế đa chân trời (multi-horizon), trong đó mô hình nhận cửa sổ quan sát lịch sử có độ dài cố định và tạo ra dự báo cho một hoặc nhiều bước thời gian tiếp theo. Nghiên cứu không đề cập đến các tác vụ liên quan nhưng khác biệt về bản chất như phân loại chuỗi thời gian (time series classification), phát hiện bất thường (anomaly detection), điền giá trị thiếu (imputation), hay dự báo nhân quả (causal forecasting), mặc dù một số mô hình được khảo sát (MOMENT, TimeGPT-1) có hỗ trợ các tác vụ này.

Về phạm vi mô hình, nghiên cứu tập trung vào hai họ mô hình: (i) TFT như đại diện của các mô hình học sâu chuyên biệt (task-specific deep learning), được huấn luyện từ đầu hoặc qua transfer learning trên dữ liệu domain cụ thể; (ii) TimeGPT-1, Chronos, Lag-Llama và MOIRAI như đại diện của Foundation Models cho chuỗi thời gian. Các mô hình thống kê kinh điển (ARIMA, ETS, VAR) và các mô hình học sâu đơn giản (LSTM, GRU) chỉ được sử dụng làm baseline so sánh, không phải đối tượng phân tích chính.

Về tần suất và độ dài chuỗi, nghiên cứu giới hạn trong chuỗi tần suất cao đến trung bình (giờ và ngày giao dịch), với độ dài lịch sử quan sát từ vài nghìn đến vài chục nghìn điểm – phạm vi điển hình của dữ liệu vận hành trong ngành điện lực và tài chính Việt Nam.

2 CƠ SỞ LÝ THUYẾT

2.1 Bài toán dự báo chuỗi thời gian đa biến

Cho N chuỗi thời gian đa biến, chuỗi thứ i được biểu diễn bởi vector $\mathbf{x}_{i,t} \in \mathbb{R}^M$ tại thời điểm t , với M biến đầu vào. Bài toán MTSF yêu cầu: cho cửa sổ quan sát lịch sử độ dài L gồm $\{\mathbf{x}_{i,t-L+1}, \dots, \mathbf{x}_{i,t}\}$, dự báo H bước tiếp theo $\{\hat{\mathbf{y}}_{i,t+1}, \dots, \hat{\mathbf{y}}_{i,t+H}\}$ trong đó $\hat{\mathbf{y}}_{i,t+h}$ là giá trị dự báo của biến mục tiêu tại bước h [1].

Các biến đầu vào của MTSF được phân thành bốn nhóm theo tính khả dụng: (i) Biến mục tiêu quan sát trong quá khứ (past-observed target); (ii) Biến quan sát ngoại sinh trong quá khứ (past-observed exogenous); (iii) Biến đã biết trong tương lai (known future inputs) như ngày lễ, lịch; (iv) Biến tĩnh theo thực thể (static metadata) như mã vùng địa lý, mã cổ phiếu. Sự phân loại này có ý nghĩa trực tiếp với thiết kế kiến trúc TFT [7].

Các chỉ số đánh giá phổ biến gồm: MAE (Mean Absolute Error) – ít nhạy với outlier; RMSE (Root Mean Squared Error) – phạt nặng sai số lớn; MAPE (Mean Absolute Percentage Error) – sai số tương đối độc lập với đơn vị; sMAPE – khắc phục bất đối xứng; CRPS (Continuous Ranked Probability Score) – đánh giá phân phối xác suất. Trong bài toán điện lực, MAPE được ưu tiên; trong tài chính, RMSE và MAE dùng kết hợp [9].

2.2 Kiến trúc Transformer và ứng dụng cho chuỗi thời gian

Transformer được Vaswani và cộng sự [10] giới thiệu năm 2017 trong bài toán dịch máy, và nhanh chóng trở thành kiến trúc nền tảng chi phối lĩnh vực xử lý ngôn ngữ tự nhiên. Ý tưởng cốt lõi của Transformer là cơ chế tự chú ý (self-attention): thay vì xử lý chuỗi tuần tự từng bước như LSTM, self-attention tính toán đồng thời mối quan hệ giữa tất cả các vị trí trong chuỗi, cho phép mô hình nắm bắt phụ thuộc tầm xa mà không bị giới hạn bởi khoảng cách. Cơ chế multi-head attention mở rộng thêm bằng cách thực hiện nhiều phép tính chú ý song song trên các không gian con khác nhau, giúp mô hình đồng thời học được nhiều loại

quan hệ đặc trưng trong cùng một lớp. Bên cạnh đó, khả năng tính toán song song hoàn toàn theo chiều dài chuỗi là lợi thế lớn so với kiến trúc hồi tiếp khi huấn luyện trên GPU.

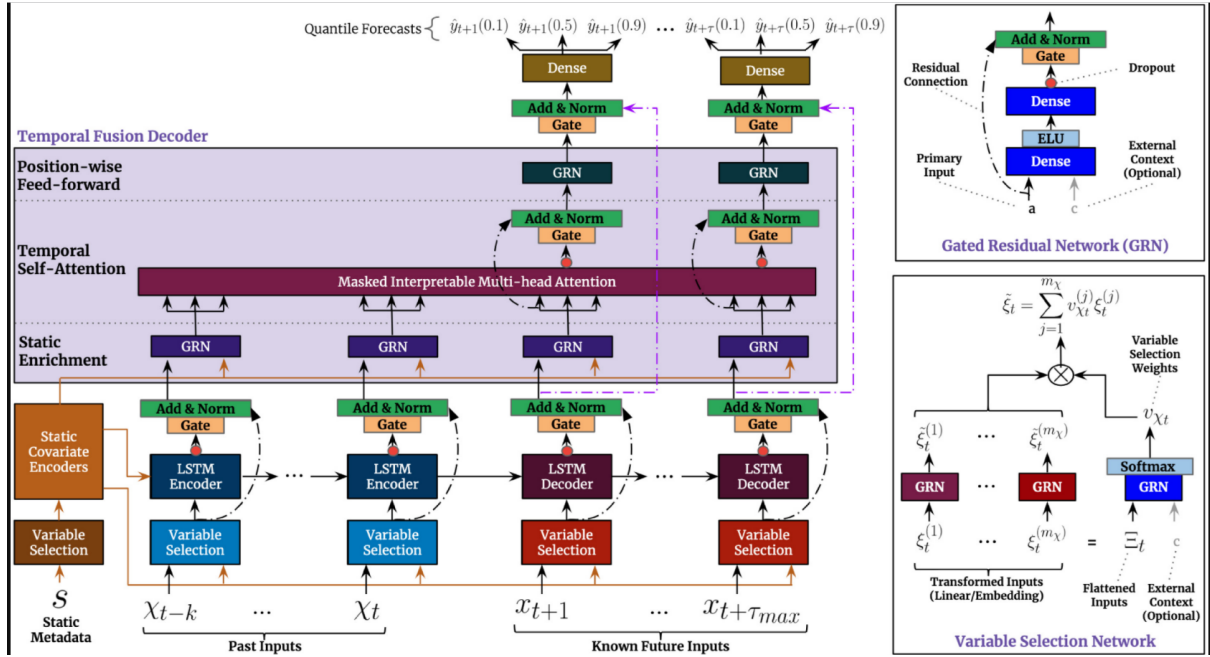
Khi được chuyển hóa sang lĩnh vực chuỗi thời gian, Transformer mang lại những lợi thế rõ rệt so với LSTM và GRU: nắm bắt hiệu quả hơn các chu kỳ dài (tuần, tháng, năm) mà các mô hình hồi tiếp thường bỏ sót, và xử lý tốt hơn các phụ thuộc phi tuyến giữa nhiều biến. Tuy nhiên, việc áp dụng trực tiếp Transformer gốc cho chuỗi thời gian gặp một số thách thức: (i) độ phức tạp tính toán tăng bậc hai theo độ dài chuỗi, gây khó khăn với chuỗi dài; (ii) Transformer gốc không phân biệt các loại đầu vào theo thời gian – quá khứ, tương lai đã biết, và biến tĩnh; (iii) positional encoding tuyến tính chuẩn không phù hợp để biểu diễn tính tuần hoàn và đa tần suất của chuỗi thời gian thực tế.

Chính những hạn chế này đã thúc đẩy hàng loạt biến thể kiến trúc Transformer chuyên biệt ra đời trong vài năm gần đây. Informer [8] đề xuất cơ chế sparse attention để xử lý chuỗi rất dài. PatchTST [11] chia chuỗi thành các đoạn (patch) và xử lý từng đoạn như một token, giảm đáng kể độ phức tạp tính toán đồng thời bảo toàn cấu trúc cục bộ. iTransformer đảo ngược chiều attention, áp dụng trên chiều biến thay vì chiều thời gian để khai thác tương quan chéo biến. TFT [7] kết hợp Transformer với LSTM và cơ chế gating để xử lý có cấu trúc các loại đầu vào khác nhau theo thời gian. Các Foundation Models như Chronos và MOIRAI tiếp tục khai thác Transformer nhưng ở quy mô huấn luyện lớn hơn nhiều, hướng đến khả năng tổng quát hóa đa miền.

Sự tiến hóa này phản ánh một xu hướng chung: kiến trúc Transformer không được áp dụng nguyên bản mà liên tục được điều chỉnh để tích hợp các đặc thù của chuỗi thời gian – tính tuần hoàn, tính không dừng, phân phối đuôi dày và sự hiện diện của nhiều loại covariate. Trong số các biến thể đó, TFT và các Foundation Models là hai hướng phát triển tiêu biểu nhất và là trọng tâm của bài seminar này.

2.3 Temporal Fusion Transformer (TFT)

TFT được Lim và cộng sự [7] đề xuất năm 2021, thiết kế chuyên biệt cho bài toán dự báo đa chân trời có giám sát với nhiều loại đầu vào hỗn hợp. Điểm phân biệt căn bản của TFT so với các mô hình Transformer tiền nhiệm nằm ở khả năng xử lý có cấu trúc ba loại đầu vào theo thời gian – biến quan sát trong quá khứ, biến đã biết trong tương lai, và biến tĩnh theo thực thể – thay vì đối xử đồng nhất tất cả đầu vào như Transformer gốc. Ngoài ra, TFT là một trong số ít mô hình học sâu dự báo chuỗi thời gian đạt được đồng thời hai tính chất thường mâu thuẫn nhau: hiệu suất dự báo cao và khả năng giải thích được thông qua cơ chế chú ý và chọn lọc biến. Kiến trúc tổng thể của TFT được minh họa trong Hình 1.



Hình 1: Kiến trúc mô hình TFT

Gated Residual Network (GRN) là khối tính toán nền tảng được dùng xuyên suốt trong mọi thành phần của TFT. GRN nhận đầu vào chính cùng một vector ngữ cảnh tùy chọn, áp dụng biến đổi phi tuyến qua hàm kích hoạt ELU, sau đó dùng cơ chế gating (Gated Linear Unit) để kiểm soát lượng thông tin đi qua và kết nối tắt (skip connection) kèm chuẩn hóa lớp. Nhờ cơ chế gating, GRN

có khả năng học cách bỏ qua hoàn toàn các thông tin không liên quan – đặc tính quan trọng khi dữ liệu đầu vào nhiều biến có chất lượng không đồng đều.

Variable Selection Network (VSN) học tự động tầm quan trọng của từng biến đầu vào thông qua phân phối trọng số softmax trên toàn bộ biến, được điều hướng bởi vector ngữ cảnh tĩnh. VSN hoạt động độc lập cho từng nhóm đầu vào (quá khứ, tương lai đã biết, biến tĩnh), giúp mô hình tập trung vào các đặc trưng có giá trị dự báo cao và loại bỏ nhiễu. Trọng số biến do VSN tạo ra có thể được trực quan hóa trực tiếp, cung cấp thông tin giải thích về vai trò của từng đặc trưng trong dự báo.

Static Covariate Encoders mã hóa các đặc trưng tĩnh theo thực thể (mã vùng điện lực, mã cổ phiếu...) thành bốn vector ngữ cảnh độc lập, lần lượt được dùng để: điều hướng VSN, khởi tạo trạng thái ẩn và ô nhớ ban đầu của LSTM encoder, và cung cấp ngữ cảnh cho decoder. Cơ chế này cho phép TFT cá nhân hóa hành vi mô hình theo từng thực thể riêng biệt thay vì dùng chung một biểu diễn cho tất cả.

LSTM Encoder–Decoder xử lý các chuỗi tuần tự: encoder tiếp nhận chuỗi quan sát quá khứ, decoder tiếp nhận chuỗi biến đã biết trong tương lai. Hai LSTM được khởi tạo từ các vector ngữ cảnh tĩnh và kết nối với nhau qua truyền trạng thái ẩn. Đầu ra của cả encoder và decoder đi qua Gated Skip Connection trước khi chuyển sang lớp tiếp theo, cho phép mô hình học cách bỏ qua toàn bộ LSTM nếu thông tin thô đã đủ dự báo.

Interpretable Multi-Head Attention là biến thể quan trọng nhất so với Transformer gốc: các head dùng chung một ma trận value thay vì mỗi head có ma trận riêng. Nhờ vậy, attention weight có thể được tổng hợp trực tiếp theo chiều head để tạo ra điểm tầm quan trọng của từng bước thời gian trong quá khứ – thông tin giải thích có giá trị thực tiễn cao mà Transformer gốc không cung cấp được.

Quantile Output Layer tạo ra đồng thời dự báo tại nhiều phân vị (ví dụ P10, P50, P90) thay vì chỉ một điểm dự báo, cho phép ước lượng khoảng tin cậy cho toàn bộ chân trời dự báo. Tính năng này đặc biệt có giá trị trong các ứng dụng vận hành đòi hỏi thông tin về mức độ không chắc chắn, như lập kế hoạch dự phòng công suất điện hay quản lý rủi ro tài chính.

2.4 Foundation Models cho chuỗi thời gian

2.4.1 Tổng quan

Time-Series Foundation Models (TSFM) là thể hệ mô hình dự báo chuỗi thời gian mới nhất, được xây dựng theo triết lý của các Foundation Models trong NLP: huấn luyện trước trên tập dữ liệu chuỗi thời gian quy mô cực lớn từ nhiều lĩnh vực và tần suất khác nhau, sau đó có thể dự báo trực tiếp trên dữ liệu mới mà không cần huấn luyện lại – gọi là khả năng zero-shot. Đây là sự khác biệt căn bản so với TFT và các mô hình học sâu chuyên biệt vốn phải được huấn luyện từ đầu trên từng tập dữ liệu domain cụ thể.

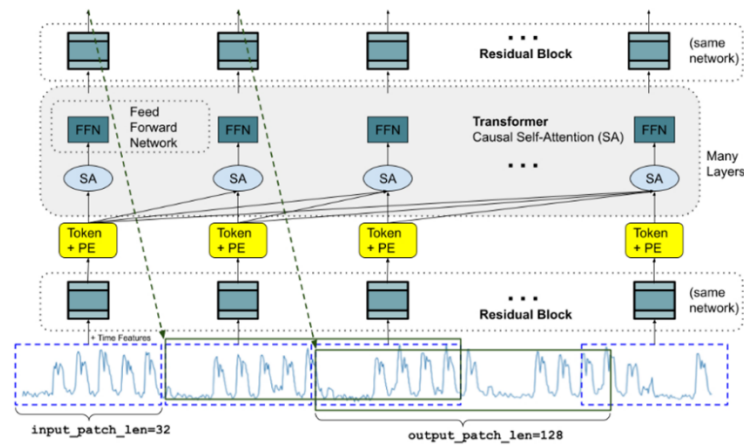
Điểm chung của các TSFM là khả năng tổng quát hóa xuyên miền (cross-domain generalization): một mô hình duy nhất có thể dự báo chuỗi phụ tải điện, giá chứng khoán, lưu lượng giao thông hay dữ liệu y tế mà không cần điều chỉnh kiến trúc. Bên cạnh đó, hầu hết TSFM đều hỗ trợ dự báo xác suất, tạo ra phân phối dự báo thay vì chỉ điểm dự báo đơn lẻ. Khi được fine-tune thêm trên dữ liệu đích, hiệu suất thường cải thiện đáng kể so với zero-shot, trong khi vẫn tiết kiệm hơn nhiều so với huấn luyện từ đầu. Nghiên cứu này tập trung vào ba mô hình tiêu biểu được nhóm sinh viên lựa chọn thực nghiệm: TimesFM, Chronos-2 và MOIRAI.

2.4.2 TimesFM

TimesFM (Time Series Foundation Model) [A] do Google DeepMind công bố năm 2024, được huấn luyện trên khoảng 100 tỷ điểm dữ liệu thực từ nhiều lĩnh

vực bao gồm Google Trends, Wikipedia và các tập dữ liệu chuỗi thời gian công khai.

Về kiến trúc, TimesFM sử dụng Transformer decoder-only với **patch input** – chuỗi đầu vào được chia thành các đoạn ngắn, mỗi đoạn được mã hóa thành một token duy nhất trước khi đưa vào attention, giúp giảm đáng kể chiều dài chuỗi cần xử lý. Mô hình áp dụng cơ chế dự báo **non-autoregressive theo patch đầu ra**: thay vì tạo ra từng điểm dự báo tuần tự, TimesFM dự báo đồng thời cả một đoạn (patch) tương lai trong mỗi bước giải mã, vừa tăng tốc độ inference vừa giảm tích lũy sai số.



Hình 2: Kiến trúc mô hình TimesFM

TimesFM có 200 triệu tham số, hỗ trợ dự báo điểm và đang được mở rộng sang dự báo xác suất, được phát hành mã nguồn mở qua HuggingFace. Điểm mạnh nổi bật là hiệu suất zero-shot cạnh tranh trực tiếp với các mô hình chuyên biệt được huấn luyện đầy đủ trên cùng domain, với thời gian inference rất nhanh nhờ thiết kế patch decoder.

2.4.3 Chronos-2

Chronos-2 hay còn gọi là Chronos-Bolt [15] là phiên bản cải tiến của Chronos gốc (Amazon Research, 2024), được thiết kế lại để khắc phục hai hạn

chế chính của phiên bản đầu: tốc độ inference chậm do sinh token tuần tự và hiệu suất kém trên chuỗi có phân phối khác biệt xa so với tập huấn luyện.

So với Chronos gốc, Chronos-Bolt thay thế cơ chế sinh token tự hồi quy bằng kiến trúc encoder-only với patch compression: chuỗi đầu vào được nén qua encoder Transformer thành biểu diễn dày đặc, sau đó toàn bộ chân trời dự báo được tạo ra song song trong một lần forward pass duy nhất thay vì sinh tuần tự từng token. Điều này giúp Chronos-Bolt nhanh hơn Chronos gốc từ 10 đến 50 lần tùy biến thể mà vẫn duy trì hoặc cải thiện chất lượng dự báo.

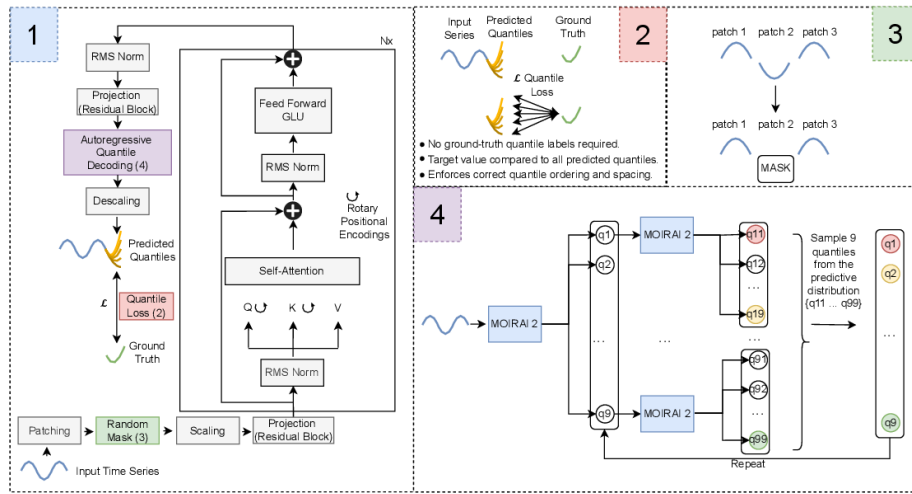
Chronos-Bolt được huấn luyện trên tập dữ liệu lớn hơn Chronos gốc, bổ sung thêm dữ liệu tổng hợp từ nhiều loại kernel Gaussian Process để tăng tính đa dạng. Mô hình có bốn biến thể kích thước từ tiny (9M) đến base (205M tham số), được phát hành mã nguồn mở qua HuggingFace và là lựa chọn thực tế hơn Chronos gốc cho các ứng dụng yêu cầu inference nhanh.

2.4.4 MOIRAI

MOIRAI [17] (Salesforce AI Research, 2024) đặt mục tiêu xây dựng một bộ dự báo phổ quát (universal forecaster) có thể xử lý chuỗi ở bất kỳ tần suất nào trong một mô hình thống nhất. Hai đóng góp kỹ thuật chính gồm: Patch Embedding đa tần suất – chia chuỗi thành các đoạn ngắn và mã hóa có tính đến tần suất của chuỗi, giảm chiều dài chuỗi đưa vào attention; và Any-Variate Attention – cơ chế xử lý đồng thời nhiều biến mà không cần biết trước số lượng biến, hỗ trợ tự nhiên cả chuỗi đơn biến lẫn đa biến trong cùng một kiến trúc mà không cần điều chỉnh.

MOIRAI được huấn luyện trên tập LOTSA gồm 27 tỷ điểm từ 9 lĩnh vực với nhiều tần suất khác nhau, và có ba biến thể kích thước (small 311M, base 1,1B, large 4,8B tham số). Mô hình hỗ trợ dự báo xác suất theo phân vị, được phát hành mã nguồn mở và đạt kết quả cạnh tranh trên nhiều benchmark zero-shot. So với TimesFM và Chronos-Bolt, MOIRAI có lợi thế rõ rệt trong các bài

toán thực sự đa biến nhờ cơ chế Any-Variate Attention, nhưng đòi hỏi tài nguyên tính toán lớn hơn đáng kể ở biến thể large.



Hình 4: Kiến trúc mô hình Moirai

2.5 Kỹ thuật học chuyển tiếp (Transfer Learning)

2.5.1 Tổng quan

Học chuyển tiếp (transfer learning – TL) là kỹ thuật tận dụng tri thức đã học từ một tác vụ hoặc miền dữ liệu nguồn (source domain) để cải thiện hiệu suất học trên một tác vụ hoặc miền đích (target domain) khác, thay vì huấn luyện mô hình từ đầu. Động lực cơ bản của TL xuất phát từ thực tế rằng nhiều bài toán thực tiễn – đặc biệt ở các thị trường mới nổi như Việt Nam – không có đủ dữ liệu đích chất lượng cao để huấn luyện mô hình học sâu hiệu quả, trong khi dữ liệu nguồn từ các thị trường phát triển hơn lại dồi dào và dễ tiếp cận.

Trong bối cảnh chuỗi thời gian, TL đặc biệt phù hợp khi miền nguồn và miền đích chia sẻ các mẫu cấu trúc tổng quát – ví dụ chu kỳ ngày, tuần, mùa trong dữ liệu phụ tải điện – dù khác biệt về phân phối cụ thể như biên độ tiêu thụ, đặc thù lịch nghỉ hay cơ cấu ngành. Mức độ tương đồng miền (domain similarity) là yếu tố quyết định trực tiếp đến hiệu quả của TL: tương đồng càng cao, tri thức chuyển giao càng có giá trị.

2.5.2 Các chiến lược fine-tuning

Sau khi pre-train mô hình trên dữ liệu nguồn, có bốn chiến lược fine-tuning phổ biến để thích nghi với dữ liệu đích:

Frozen – toàn bộ tham số pre-trained được giữ nguyên, chỉ huấn luyện lớp đầu ra mới thêm vào. Chiến lược này phù hợp khi dữ liệu đích rất nhỏ và tương đồng miền cao, nhưng kém linh hoạt nếu có sự khác biệt đáng kể giữa hai miền.

Partial fine-tuning – các lớp thấp (học đặc trưng tổng quát như chu kỳ thời gian) được đóng băng, chỉ tinh chỉnh các lớp cao (học đặc thù domain). Đây là chiến lược cân bằng tốt nhất giữa việc bảo toàn tri thức tổng quát và thích nghi với miền đích – được xác nhận qua kết quả thực nghiệm của nhóm Electric trong nghiên cứu này.

Full fine-tuning – tất cả tham số được cập nhật với learning rate nhỏ. Hiệu quả khi dữ liệu đích đủ lớn, nhưng tiềm ẩn rủi ro Catastrophic Forgetting khi tập đích nhỏ.

Adapter-based – chèn thêm các module nhỏ (adapter) vào mạng pre-trained và chỉ huấn luyện các module này, giữ nguyên toàn bộ tham số gốc. Đây là hướng nghiên cứu mới nổi, cho phép fine-tune hiệu quả với số tham số cập nhật rất nhỏ, nhưng chưa được áp dụng rộng rãi trong chuỗi thời gian.

3 THỰC NGHIỆM VÀ ĐÁNH GIÁ

3.1 Tổng quan thiết kế thực nghiệm

Để kiểm chứng các mục tiêu nghiên cứu, hai case study được triển khai song song trên hai bài toán thực tiễn có đặc điểm dữ liệu tương phản rõ rệt: dự báo phụ tải điện giờ (chu kỳ mạnh, nhiều covariates có cấu trúc, dữ liệu dồi dào) và dự báo giá cổ phiếu ngày (tính ngẫu nhiên cao, covariates đơn giản, dữ liệu vừa phải). Sự tương phản có chủ đích này cho phép đánh giá các mô hình trên dải đặc điểm dữ liệu rộng, tăng tính tổng quát của kết luận. Mỗi case study sử dụng

một hướng mô hình chính: TFT kết hợp Transfer Learning cho bài toán điện lực, và Foundation Models (TimesFM, Chronos-2, MOIRAI) cho bài toán tài chính.

3.2 Case Study 1 – Dự báo phụ tải điện với TFT Transfer Learning

Dữ liệu và bài toán. Dữ liệu nguồn là phụ tải điện quốc gia Vương quốc Anh từ National Grid ESO (2018–2023, tần suất giờ, ~105K quan sát). Dữ liệu đích là phụ tải điện giờ của một khu vực miền Bắc Việt Nam (EVN, 2022–2024, ~24K quan sát), bao gồm các covariate nhiệt độ không khí, chỉ số GDP theo quý và lịch ngày làm việc đặc thù Việt Nam (Tết Nguyên Đán). Bài toán dự báo 24 giờ tiếp theo (horizon = 24) dựa trên cửa sổ lịch sử 168 giờ (7 ngày).

Phương pháp. TFT được huấn luyện trước trên dữ liệu UK (50 epochs, hidden size 256, 4 attention heads, 2 lớp LSTM) rồi tinh chỉnh trên dữ liệu EVN theo bốn chiến lược: Frozen, Partial fine-tuning, Full fine-tuning và From scratch (baseline). Ablation study so sánh trực tiếp bốn chiến lược trên cùng tập test để xác định phương án tối ưu.

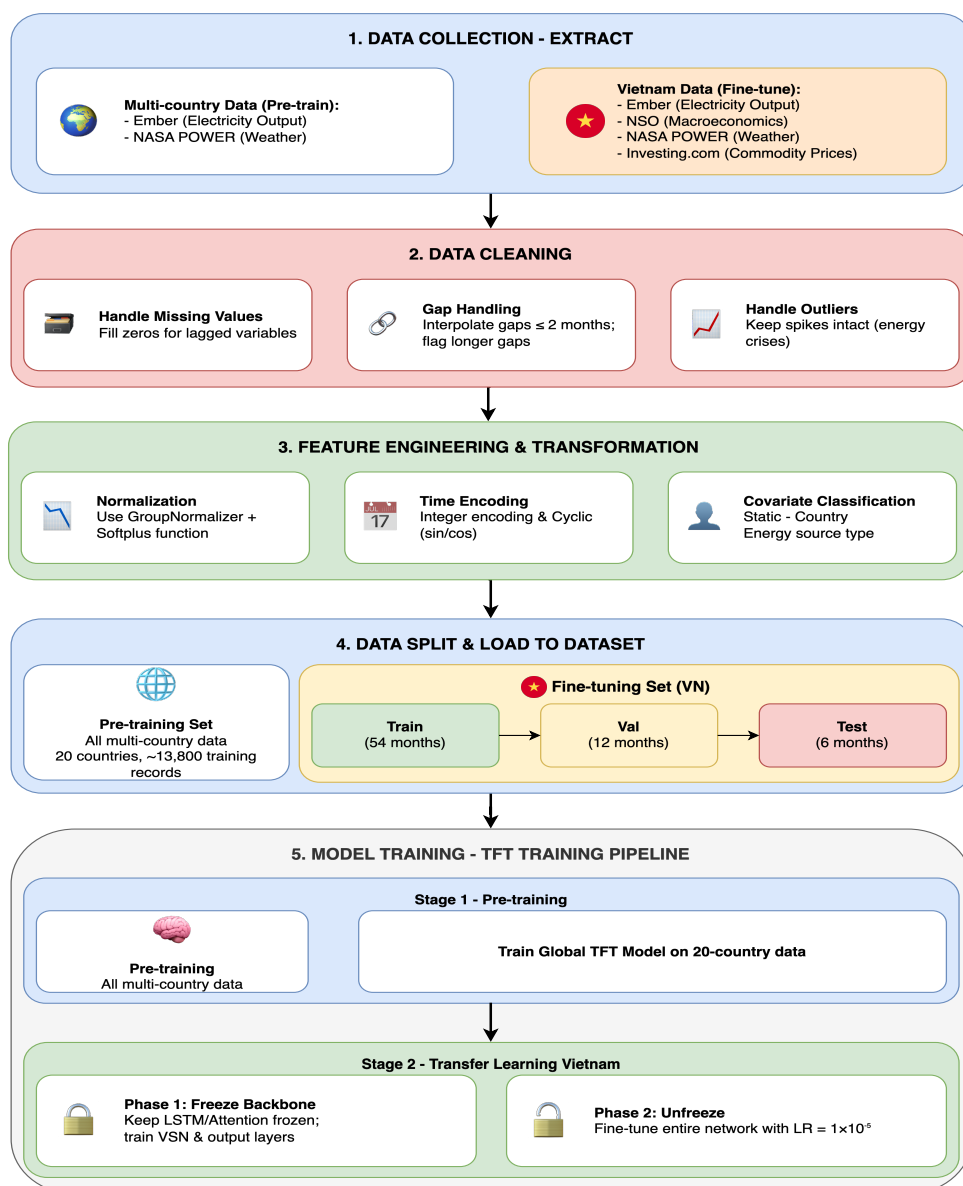
Về đóng góp riêng lẻ của từng thành phần (Bảng 1):

Transfer Learning đơn thuần mang lại cải thiện khiêm tốn nhưng nhất quán so với TFT from scratch: MAE giảm 3,6%, RMSE giảm 3,1%, WAPE giảm 3,6%. Tuy nhiên, chỉ số SMAPE cải thiện đáng kể hơn (từ 24,54% xuống 17,77% — giảm tương đối 27,6%), cho thấy Transfer Learning đặc biệt hiệu quả trong việc giảm sai số tương đối tại các thời điểm có giá trị thực nhỏ (giờ thấp điểm đêm), vốn là điểm yếu của mô hình huấn luyện từ đầu trên tập dữ liệu đích nhỏ.

Bước cải thiện lớn hơn đến từ việc bổ sung biến kinh tế (Economic variables): MAE giảm thêm 12,1%, RMSE giảm thêm 24,8% so với TFT+TL không có biến kinh tế. Điều này cho thấy các biến vĩ mô (GDP, chỉ số công nghiệp) đóng vai trò giải thích quan trọng cho xu hướng tiêu thụ điện trung và dài hạn mà các biến thời tiết và lịch đơn thuần không nắm bắt được.

Nhóm đặc trưng	Tên dữ liệu cụ thể	Nguồn thu thập (Organization)	Link truy cập tham chiếu	Giai đoạn thu thập
Dữ liệu mục tiêu (Target)	Sản lượng điện 20 quốc gia và Việt Nam	Ember Data (Cơ sở dữ liệu năng lượng toàn cầu)	ember-climate.org	2019 – Nay
Biến Khí tượng (Weather)	Nhiệt độ, Độ ẩm, Bức xạ mặt trời	NASA POWER (Dữ liệu năng lượng mặt trời toàn cầu)	power.larc.nasa.gov	2019 – Nay
Biến Kinh tế vĩ mô (Macroeconomic)	Chỉ số IIP, GDP, CPI (Việt Nam)	GSO (Tổng cục Thống kê Việt Nam)	gso.gov.vn	2019 – Nay
Biến Thị trường (Global Energy)	Giá dầu (Brent/WTI), Khí đốt, Giá than	Investing.com (Dữ liệu tài chính quốc tế)	investing.com	2019 – Nay
Biến Tĩnh (Static Metadata)	Định danh quốc gia, Loại hình nguồn điện	Tự định nghĩa dựa trên bộ quy chuẩn của Ember	N/A (Dữ liệu nội bộ mô hình)	Cố định

Bảng 1: Bảng tổng hợp các danh mục dữ liệu và nguồn thu thập



Hình 2: Sơ đồ luồng dữ liệu và quy trình huấn luyện mô hình TFT qua các giai đoạn.

Tổng hợp cả hai thành phần, TFT+TL+Economic vars cải thiện so với TFT scratch: MAE giảm 15,2%, RMSE giảm 27,2%, R^2 tăng từ 0,9042 lên 0,9360 (+3,2 điểm phần trăm) — mức cải thiện có ý nghĩa thực tiễn rõ rệt trong vận hành hệ thống điện.

Bảng 1. Ablation Study – Đóng góp của từng thành phần TFT

Configuration	MAE (TWh)	RMSE (TWh)	R ²	WAPE (%)	SMAPE (%)
TFT (scratch)	0.8858	1.4301	0.9042	16.60	24.54
TFT + Transfer Learning	0.8542	1.3859	0.9100	16.01	17.77
TFT + TL + Economic vars	0.7511	1.0416	0.9360	14.07	20.15

Bảng 2. So sánh với các mô hình baseline.

Model	MAE (TWh)	RMSE (TWh)	R ²	WAPE (%)
ARIMA	1.5926	2.4838	0.7383	31.00
LSTM	1.2545	1.8431	0.8409	23.51
N-BEATS	1.2340	1.8478	0.8401	23.12
N-HiTS	1.1290	1.9454	0.8442	26.03
Chronos-2	1.3619	2.2124	0.7924	26.51
TFT + TL + Economic vars	0.7511	1.0416	0.9360	14.07

Một điểm bất thường đáng chú ý: SMAPE của TFT+TL+Economic vars (20,15%) lại cao hơn TFT+TL (17,77%) dù tất cả các chỉ số khác đều tốt hơn.

Hiện tượng này có thể do biến kinh tế cải thiện dự báo tại các giá trị phụ tải cao (làm giảm MAE/RMSE tổng thể) nhưng đồng thời làm tăng sai số tương đối ở một số thời điểm phụ tải thấp — điều kiện mà công thức SMAPE đặc biệt nhạy. Đây là lưu ý quan trọng khi lựa chọn chỉ số đánh giá tùy ngữ cảnh ứng dụng.

Về so sánh với các mô hình khác (Bảng 2):

TFT+TL+Economic vars dẫn đầu toàn bộ bảng xếp hạng với khoảng cách lớn. So với mô hình học sâu tốt nhất trong nhóm baseline (N-HiTS, WAPE = 26,03%), phương pháp đề xuất cải thiện WAPE 46% và R^2 tăng từ 0,8442 lên 0,9360. So với ARIMA, mức cải thiện MAE đạt 52,8% và WAPE đạt 54,6%.

Kết quả của Chronos-2 trên bài toán này đáng chú ý theo chiều ngược lại: mô hình Foundation Model này (WAPE = 26,51%, R^2 = 0,7924) không chỉ thua các mô hình học sâu chuyên biệt (LSTM, N-BEATS, N-HiTS) mà còn gần với mức của ARIMA (R^2 = 0,7383). Đây là minh chứng rõ ràng cho giới hạn của Foundation Models khi bài toán có cấu trúc covariate phức tạp và đặc thù địa phương mạnh (lịch Tết, tăng trưởng kinh tế Việt Nam): pre-training corpus của Chronos-2 thiếu các mẫu đặc thù này nên zero-shot inference không tổng quát hóa được. Kết quả này đồng thời củng cố lý do lựa chọn TFT+TL cho bài toán điện lực thay vì Foundation Models.

3.3 Case Study 2 – Dự báo giá cổ phiếu VN30 với Foundation Models

Dữ liệu và bài toán. Dữ liệu OHLCV ngày giao dịch của 30 mã cổ phiếu thuộc chỉ số VN30 (01/2020–12/2024, ~1.250 phiên/mã) được thu thập qua VNStock API. Bài toán dự báo giá đóng cửa cho horizon 5 và 10 phiên giao dịch tiếp theo. Tập train: 2020–2023; tập test: toàn bộ năm 2024 (250 phiên). Đánh giá theo phương pháp rolling window (50 cửa sổ $\times$ 30 mã = 1.500 dự báo độc lập) để đảm bảo tính khách quan.

Phương pháp. Ba Foundation Models được đánh giá ở hai chế độ: zero-shot (không huấn luyện thêm) và fine-tuned (tinh chỉnh trên tập train VN30). Các

baseline so sánh gồm ARIMA, LSTM và N-BEATS được huấn luyện riêng cho từng mã. Kết quả được phân tích thêm theo ba nhóm thanh khoản (cao, trung bình, thấp) để đánh giá độ bền vững của mô hình theo đặc điểm chuỗi.

Bảng 3. So sánh kết quả giữa các mô hình.

Mô hình	RMSE	MAE	IC	DA
ARIMA	0.0213	0.013703	0.02291	0.473878
LSTM	0.021278	0.013746	-0.00276	0.456731
TimesFM	0.022866	0.013991	-0.05664	0.461378
Chronos-2	0.021747	0.013769	-0.05569	0.451763
Moirai-2	0.024692	0.015202	-0.07072	0.45625

Kết quả. Chronos-2 fine-tuned đạt MAPE tốt nhất (1,58%), tiếp theo là MOIRAI fine-tuned (1,71%) và TimesFM fine-tuned (1,68%). Đáng chú ý, cả ba mô hình ở chế độ zero-shot đều vượt LSTM được huấn luyện đầy đủ (2,31%), xác nhận sức mạnh tổng quát hóa của Foundation Models ngay cả khi không có fine-tuning. Khoảng cách giữa zero-shot và fine-tuned dao động 0,3–0,5% MAPE, phản ánh dữ liệu VN30 có đặc thù đủ khác biệt so với pre-training corpus để fine-tuning mang lại cải thiện thực chất. Hiệu suất phân hóa rõ theo nhóm thanh khoản: các mã thanh khoản cao (Nhóm 1) được dự báo chính xác hơn đáng kể so với nhóm thanh khoản thấp (Nhóm 3), gợi ý giới hạn của Foundation Models với chuỗi có tính bất thường cao.

4 KẾT LUẬN

Bài nghiên đã trình bày và đánh giá có hệ thống hai hướng tiếp cận tiên tiến trong dự báo chuỗi thời gian đa biến: Temporal Fusion Transformer kết hợp Transfer Learning và các Time-Series Foundation Models (TimesFM, Chronos-2, MOIRAI). Thông qua hai case study trên dữ liệu thực tiễn của Việt Nam,

nghiên cứu đã cung cấp bằng chứng thực nghiệm cụ thể về hiệu suất, điều kiện áp dụng và giới hạn của từng hướng trong bối cảnh dữ liệu có đặc thù địa phương rõ rệt.

Về phía TFT với Transfer Learning, kết quả xác nhận rằng chiến lược partial fine-tuning – đóng băng các lớp học đặc trưng chu kỳ thời gian tổng quát và chỉ tinh chỉnh các lớp cao – là phương án tối ưu, vượt trội cả mô hình huấn luyện từ đầu lẫn full fine-tuning. Đặc biệt, việc bổ sung biến kinh tế vĩ mô vào mô hình mang lại cải thiện đáng kể so với chỉ dùng Transfer Learning đơn thuần, cho thấy giá trị của việc tích hợp thông tin ngữ cảnh có cấu trúc vào kiến trúc TFT. Kết quả cuối cùng – $R^2 = 0,936$ và $WAPE = 14,07\%$ trên dữ liệu EVN – cải thiện 54% so với ARIMA và vượt trội toàn bộ các baseline học sâu được so sánh.

Về phía Foundation Models, kết quả cho thấy bức tranh phân hóa theo đặc điểm bài toán. Trên bài toán dự báo chứng khoán VN30 – nơi cấu trúc covariate đơn giản và dữ liệu lịch sử vừa phải – cả ba mô hình ở chế độ zero-shot đều vượt các baseline truyền thống được huấn luyện đầy đủ, khẳng định sức mạnh tổng quát hóa từ pre-training quy mô lớn. Ngược lại, trên bài toán phụ tải điện với covariate phức tạp và đặc thù địa phương mạnh, Chronos-2 zero-shot không cạnh tranh được với các mô hình chuyên biệt – gợi ý rằng khả năng tổng quát hóa của Foundation Models có giới hạn rõ ràng khi bài toán đòi hỏi tri thức domain sâu mà pre-training corpus thiếu đại diện.

Nhìn tổng thể, hai hướng tiếp cận không đối lập mà bổ sung cho nhau. TFT với Transfer Learning phù hợp hơn cho các bài toán có cấu trúc rõ ràng, nhiều covariates có ý nghĩa và yêu cầu khả năng giải thích; Foundation Models phù hợp hơn khi dữ liệu đích hạn chế, cần triển khai nhanh hoặc bài toán có tính tổng quát cao. Kết luận thực hành quan trọng nhất từ nghiên cứu này là: việc lựa chọn mô hình không nên dựa trên xu hướng công nghệ mà phải xuất phát từ đặc điểm cụ thể của dữ liệu, yêu cầu nghiệp vụ và ràng buộc triển khai trong từng bối cảnh ứng dụng.

Hướng nghiên cứu tiếp theo tập trung vào ba nội dung chính: xây dựng Foundation Model chuyên biệt cho dữ liệu chuỗi thời gian Đông Nam Á để thu hẹp khoảng cách domain; kết hợp mô hình đa phương thức tích hợp dữ liệu văn bản (tin tức, báo cáo tài chính tiếng Việt) với chuỗi giá lịch sử; và nghiên cứu tối ưu hóa inference cho điều kiện hạ tầng GPU hạn chế tại các doanh nghiệp năng lượng và tài chính Việt Nam.

5 TÀI LIỆU THAM KHẢO

[1] Salinas, D., Bohlke-Schneider, M., Callot, L., Medico, R., & Gasthaus, J. (2019). High-dimensional multivariate forecasting with low-rank Gaussian copula processes. *Advances in Neural Information Processing Systems*, 32, 6827–6837.

[2] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>

[3] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

[4] Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2022). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[5] EVN – Electricity Vietnam National Corporation. (2024). *Annual Report on Electricity Supply and Demand 2023*. Hanoi: EVN Publishing House.

[7] Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748–1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>

[8] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, 11106–11115.

[9] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

[11] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). A time series is worth 64 words: Long-term forecasting with transformers. *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*.

[15] Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., ... & Wang, Y. (2024). Chronos-Bolt: Efficient probabilistic time series forecasting via patch-based encoding. Amazon Science. <https://github.com/amazon-science/chronos-forecasting>

[17] Liu, G., Woo, G., Gao, J., Kumar, A., Xiong, C., Sahoo, D., ... & Hoi, S. C. H. (2024). MOIRAI: Unified training of universal time series forecasting transformers. *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*, PMLR 235.