TRƯỜNG ĐẠI HỌC QUY NHƠN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO TIỂU LUẬN

HỌC PHẦN: MÔ HÌNH DỰ BÁO CHUỖI THỜI GIAN

ĐỀ TÀI: DỰ BÁO TIÊU THỤ ĐIỆN NĂNG BẰNG MÔ HÌNH KẾT HỢP XGBOOST VÀ LSTM TRÊN CHUỖI THỜI GIAN

Giảng viên hướng dẫn:	TS. LÊ XUÂN VIỆT
Lớp học phần:	1050363
Sinh viên thực hiện:	HÀ NHẬT ĐOAN
Mã sinh viên:	4554100017
Khoa:	CÔNG NGHỆ THÔNG TIN
Ngành và khóa:	TRÍ TUỆ NHÂN TẠO K45

LỜI CẢM ƠN

Kính gửi Thầy Lê Xuân Việt,

Lời đầu tiên, em xin gửi đến Thầy lời cảm ơn chân thành và sâu sắc nhất vì những kiến thức và sự tận tâm mà Thầy đã dành cho chúng em trong suốt học phần mô hình dự báo chuỗi thời gian vừa qua.

Nhờ sự hướng dẫn tận tình và phương pháp giảng dạy khoa học, dễ hiểu của Thầy, em đã có cơ hội tiếp cận và nắm vững những kiến thức nền tảng cũng như các kỹ thuật phân tích và dự báo chuỗi thời gian quan trọng. Từ những khái niệm cơ bản đến các mô hình phức tạp, Thầy đã truyền đạt một cách hệ thống và logic, giúp chúng em hiểu rõ bản chất của vấn đề và cách áp dụng chúng vào thực tế.

Không chỉ dừng lại ở việc truyền đạt kiến thức chuyên môn, Thầy còn khơi gợi niềm hứng thú và đam mê nghiên cứu trong em. Những ví dụ thực tế và những chia sẻ kinh nghiệm quý báu của Thầy đã giúp em nhận thấy rõ hơn vai trò và ứng dụng rộng rãi của mô hình dự báo chuỗi thời gian trong nhiều lĩnh vực khác nhau. Điều này đã tạo động lực lớn để em tiếp tục tìm hiểu và khám phá sâu hơn về lĩnh vực này.

Em xin trân trọng cảm ơn sự nhiệt tình, tâm huyết và những nỗ lực không ngừng nghỉ của Thầy trong suốt quá trình giảng dạy. Thầy không chỉ là một người thầy giỏi mà còn là một người truyền cảm hứng lớn đối với em.

Cuối cùng, em xin kính chúc Thầy luôn mạnh khỏe, hạnh phúc và thành công trên con đường sự nghiệp. Em hy vọng sẽ có cơ hội được học hỏi thêm nhiều điều từ Thầy trong tương lai.

Trân trọng,

Hà Nhật Đoan TTNT K45

Nội dung

Tóm tắt.....5
1 Giới thiệu.....5
2 Cơ sở lý thuyết.....6
3 Phương pháp chính.....7
4 Thực nghiệm và đánh giá.....13
5 Kết luận.....20
6 Một số công trình liên quan.....21
Tài liệu tham khảo.....22

XGBoost LSTM: Dự báo tiêu thụ điện năng bằng mô hình kết hợp XGBoost và LSTM trên chuỗi thời gian

Nhat Doan Ha[1], Viet Xuan Le[1]
[1]Khoa CNTT, Trường đại học Quy Nhơn, Việt Nam
Tác giả liên hệ: Nhat Doan Ha (hanhatdoan7889@gmail.com) và Viet Xuan Le

Tóm tắt

Dự báo tiêu thụ điện năng là một bài toán then chốt trong lĩnh vực năng lượng, đóng vai trò quan trọng trong việc lập kế hoạch vận hành hệ thống điện và tối ưu hóa nguồn lực. Tuy nhiên, dữ liệu tiêu thụ điện có tính biến động và phụ thuộc thời gian cao, gây khó khăn cho các phương pháp dự báo truyền thống. Trong tiểu luận này, tôi đề xuất một mô hình kết hợp giữa XGBoost và LSTM nhằm khai thác hiệu quả các đặc trưng phi tuyến và mối quan hệ phụ thuộc chuỗi thời gian trong dữ liệu. XGBoost được sử dụng để trích xuất đặc trưng nâng cao từ các giá trị tiêu thụ điện lịch sử, sau đó LSTM xử lý các đặc trưng này để dự báo giá trị trong tương lai. Mô hình được áp dụng trên tập dữ liệu tiêu thụ điện năng theo giờ, và kết quả cho thấy phương pháp kết hợp này mang lại độ chính xác cao, ổn định hơn so với các mô hình đơn lẻ. Kết quả dự báo 500 bước tiếp theo cũng phản ánh rõ xu hướng và chu kỳ trong tiêu thụ điện, chứng minh tính khả thi của phương pháp trong ứng dụng thực tiễn.

Từ khóa: Dự báo chuỗi thời gian, tiêu thụ điện năng, XGBoost, LSTM, học máy, mô hình lai, đặc trưng thời gian.

1 Giới thiệu

Trong bối cảnh nhu cầu năng lượng ngày càng tăng cùng với sự biến động khó lường của hệ thống điện, việc dự báo chính xác mức tiêu thụ điện năng trở thành một bài toán quan trọng nhằm hỗ trợ điều độ lưới điện, tối ưu hóa vận hành và giảm thiểu

chi phí vận hành hệ thống điện. Tuy nhiên, dữ liệu tiêu thụ điện năng thường mang tính chu kỳ, phi tuyến và dễ bị ảnh hưởng bởi các yếu tố khí hậu, kinh tế - xã hội, khiến cho việc dự báo trở nên đầy thách thức đối với các mô hình thống kê truyền thống như ARIMA hay Holt-Winters.

Sự phát triển nhanh chóng của các kỹ thuật học máy (Machine Learning) và học sâu (Deep Learning) đã mở ra nhiều hướng tiếp cận mới cho bài toán này. Trong đó, mô hình Long Short-Term Memory (LSTM), một biến thể của mạng nơ-ron hồi tiếp (RNN), đã chứng minh khả năng ghi nhớ thông tin trong dài hạn và mô hình hóa các chuỗi thời gian phức tạp với mối quan hệ phụ thuộc theo thời gian hiệu quả [1]. Bên cạnh đó, mô hình Extreme Gradient Boosting (XGBoost) đã cho thấy hiệu quả vượt trội trong việc khai thác các đặc trưng phi tuyến của dữ liệu nhờ kỹ thuật boosting và tối ưu hóa dựa trên cây quyết định [2].

Việc kết hợp XGBoost và LSTM là một hướng đi tiềm năng nhằm tận dụng lợi thế của cả hai mô hình: LSTM đảm nhiệm vai trò học đặc trưng theo chuỗi thời gian trong khi XGBoost học các quan hệ phi tuyến giữa các đặc trưng đầu ra và các thông số thời gian đã được trích xuất. Cách tiếp cận kết hợp này đã được chứng minh hiệu quả trong nhiều bài toán dự báo thực tế như dự báo tài chính, thời tiết và đặc biệt là dự báo năng lượng.

Lấy cảm hứng từ cấu trúc phân rã và tái cấu trúc đa chiều được đề xuất trong nghiên cứu *iTransformer: Inverted Transformers Are Effective for Time Series Forecasting* [3], đề tài này xây dựng một mô hình kết hợp giữa XGBoost và LSTM để dự báo tiêu thụ điện năng trong tương lai dựa trên dữ liệu chuỗi thời gian. Nghiên cứu này nhằm kiểm chứng hiệu quả của mô hình kết hợp thông qua các tiêu chí đánh giá phổ biến như RMSE, MAE và MAPE, đồng thời so sánh với các mô hình đơn truyền thống.

2 Cơ sở lý thuyết

Dự báo tiêu thụ điện năng không đơn thuần là một bài toán thống kê, mà còn là một thách thức phức tạp mang tính thời gian và phi tuyến cao. Đặc điểm của dữ liệu tiêu thụ điện thường là các chuỗi thời gian dài, có sự lặp lại theo mùa, những thay đổi ngẫu nhiên không lường trước, và chịu ảnh hưởng bởi nhiều yếu tố ngoại sinh như thời tiết, hành vi người tiêu dùng, hay thậm chí là các sự kiện xã hội. Chính vì vậy, để xây dựng một mô hình dự báo hiệu quả, ta cần hiểu rõ bản chất của chuỗi thời gian cũng như lựa chọn được các mô hình học máy phù hợp.

Trong số các mô hình học sâu, LSTM (Long Short-Term Memory) nổi lên như một công cụ mạnh mẽ giúp ghi nhớ những thông tin dài hạn trong chuỗi dữ liệu. Không giống như mạng nơ-ron hồi tiếp truyền thống (RNN), LSTM được thiết kế với các cổng đặc biệt nhằm tránh tình trạng mất dần thông tin trong quá trình lan truyền ngược, từ đó học được mối quan hệ dài hạn giữa các bước thời gian. Điều này đặc biệt hữu ích trong bài toán dự báo tiêu thụ điện năng, nơi mà mức tiêu thụ hiện tại có thể phụ thuộc vào cả xu hướng dài hạn và các biến động mùa vụ trong quá khứ. [1]

Tuy nhiên, LSTM dù mạnh mẽ, vẫn có những giới hạn nhất định. Một trong số đó là việc mô hình này hoạt động như một "hộp đen", khó diễn giải và có thể gặp khó khăn trong việc nắm bắt các quan hệ phi tuyến phức tạp khi đặc trưng chưa được xử lý kỹ. Đây là lúc XGBoost (Extreme Gradient Boosting) trở thành một "đồng đội" đáng giá. Là một thuật toán boosting mạnh mẽ, XGBoost nổi bật nhờ khả năng xử lý hiệu quả dữ liệu có tính phi tuyến, kiểm soát overfitting tốt và dễ dàng diễn giải mô hình thông qua tầm quan trọng của đặc trưng. Khi kết hợp với LSTM, XGBoost có thể đảm nhiệm vai trò xử lý các đặc trưng tĩnh hoặc đầu ra từ LSTM để đưa ra dự đoán cuối cùng một cách chính xác hơn. [2]

Việc kết hợp LSTM và XGBoost không chỉ là ghép hai mô hình mạnh lại với nhau, mà là sự phối hợp giữa hai triết lý học máy: một bên tận dụng khả năng ghi nhớ và mô hình hóa thời gian, một bên tận dụng khả năng hồi quy phi tuyến mạnh mẽ. Một số phương pháp đã áp dụng chiến lược kết hợp này như: huấn luyện LSTM để trích xuất đặc trưng ẩn từ chuỗi dữ liệu, sau đó đưa các đặc trưng này làm đầu vào cho XGBoost; hoặc huấn luyện song song và tổng hợp đầu ra từ cả hai mô hình. Các nghiên cứu gần đây đã cho thấy, mô hình kết hợp này có thể vượt trội so với việc sử dụng từng mô hình đơn lẻ, nhất là trong bối cảnh dữ liệu tiêu thụ điện năng ngày càng lớn và phức tạp hơn. [3]

Hiểu rõ lý thuyết nền tảng phía sau mỗi mô hình là bước đầu tiên để xây dựng một hệ thống dự báo đáng tin cậy. Phần tiếp theo sẽ trình bày cụ thể cách tiếp cận mô hình hóa dữ liệu tiêu thụ điện năng bằng sự kết hợp giữa LSTM và XGBoost, từ khâu xử lý dữ liệu đến huấn luyện và đánh giá mô hình.

3 Phương pháp chính

3.1 Định nghĩa bài toán

Bài toán đặt ra là dự báo tiêu thụ điện năng trong tương lai dựa trên chuỗi dữ liệu tiêu thụ điện năng đã có trong quá khứ. Đây là một bài toán điển hình trong dự báo

chuỗi thời gian (time series forecasting), nơi dữ liệu có tính tuần tự, phụ thuộc theo thời gian và có thể chứa các yếu tố xu hướng (trend), mùa vụ (seasonality), và nhiễu (noise) [1].

Mục tiêu là xây dựng một mô hình học máy f(·), sử dụng đầu vào là một chuỗi con có độ dài cố định lấy từ dữ liệu lịch sử tiêu thụ điện năng, để dự đoán giá trị tiêu thụ điện năng tại một thời điểm trong tương lai gần.

Giả sử chuỗi dữ liệu tiêu thụ điện năng được biểu diễn là một dãy số rời rạc theo thời gian:

$$X = \{x_1, x_2, ..., x_T\}, x_t \in \mathbf{R} \qquad (1)$$

Với:

- x_t: Lượng điện tiêu thụ tại thời điểm t
- T: Tổng số điểm dữ liệu quan sát được

Ta đặt mục tiêu dự báo giá trị tiêu thụ h bước tiếp theo dựa trên n bước trước đó:

Trong đó:

- h: số bước thời gian dự báo trong tương lai (ví dụ: h=1 là dự báo cho bước tiếp theo)
- n: độ dài chuỗi đầu vào (window size)
- f: mô hình học máy, ở đây là sự kết hợp giữa **XGBoost** và **LSTM**

3.2 XGBoost

XGBoost, viết tắt của Extreme Gradient Boosting, là một thuật toán học máy dựa trên kỹ thuật boosting – nơi nhiều mô hình yếu (thường là các cây quyết định nhỏ) được kết hợp lại để tạo thành một mô hình mạnh mẽ. Ra đời với mục tiêu tối ưu hóa tốc độ huấn luyện và độ chính xác, XGBoost nhanh chóng trở thành công cụ phổ biến trong nhiều bài toán thực tế, đặc biệt là các cuộc thi trên Kaggle hay các bài toán hồi quy và phân loại có dữ liệu bảng [2].

Nguyên lý hoạt động của XGBoost xoay quanh ý tưởng cải thiện dần mô hình bằng cách thêm vào từng cây nhỏ sao cho cây mới học từ phần sai số (residual) mà mô hình trước chưa học tốt. Ở mỗi bước huấn luyện t, dự đoán được cập nhật như sau:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \qquad (2)$$

Tổng thể mô hình sẽ là:

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad \text{với } f_t \in \mathcal{F} \tag{3}$$

Trong đó F là tập hợp các hàm ánh xạ dạng cây quyết định.
XGBoost tối ưu hóa hàm mục tiêu gồm hai thành phần: hàm mất mát l, và một hàm điều chuẩn Ω giúp kiểm soát độ phức tạp của mô hình:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \tag{4}$$

Với

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

trong đó T là số lá trong cây, wj là giá trị được gán cho mỗi lá, và các hệ số γ, λ là siêu tham số để điều chỉnh mức phạt. Nhờ có cơ chế này, XGBoost không chỉ chính xác mà còn có khả năng chống overfitting tốt.

Trong bài toán dự báo tiêu thụ điện năng – một bài toán thuộc loại chuỗi thời gian – XGBoost không hoạt động trên chuỗi trực tiếp như các mô hình RNN hay LSTM. Thay vào đó, ta cần biến đổi dữ liệu chuỗi thành dạng bảng, thông qua các đặc trưng như độ trễ (lag), rolling mean, chỉ số thời gian (giờ, ngày, tháng), hoặc các biến phân loại như ngày lễ, cuối tuần. Một khi được chuẩn bị kỹ lưỡng, XGBoost có thể khai thác rất tốt các mối quan hệ phi tuyến giữa đặc trưng đầu vào và lượng tiêu thụ điện năng đầu ra.

Tuy không có bộ nhớ nội tại như LSTM để học được mối liên kết theo thời gian dài, nhưng XGBoost lại có lợi thế trong việc học từ những tín hiệu "ngắn hạn" mạnh mẽ, như xu hướng tăng đột ngột vào khung giờ cao điểm, hay giảm sâu vào cuối tuần. Trong nghiên cứu này, chúng tôi không chỉ dùng XGBoost để dự đoán trực tiếp, mà còn để trích xuất đặc trưng giúp mô hình LSTM phía sau nhận được đầu vào có chất lượng hơn, tăng cường khả năng ghi nhớ và dự báo chính xác.

Sự kết hợp XGBoost–LSTM là một hướng tiếp cận mang tính chiến lược: trong khi XGBoost học nhanh, chính xác và tổng quát từ những đặc trưng được mã hóa tốt, thì LSTM lại tận dụng khả năng xử lý dữ liệu tuần tự và ghi nhớ dài hạn để hiệu

chỉnh thêm kết quả cuối cùng. Mô hình lai này đã được nhiều nghiên cứu gần đây chứng minh là có hiệu quả cao trong bài toán dự báo chuỗi thời gian phức tạp như tiêu thụ điện năng [4].

3.3 LSTM

Long Short-Term Memory (LSTM) là một loại mạng nơ- ron hồi tiếp (RNN) được thiết kế để giải quyết vấn đề độ dốc biến mất trong quá trình huấn luyện các chuỗi dài. Không giống như các RNN truyền thống, LSTM có khả năng học và duy trì các phụ thuộc dài hạn, điều này giúp mô hình đặc biệt hiệu quả trong các tác vụ yêu cầu xử lý chuỗi như dịch máy, nhận dạng giọng nói và hiểu ngôn ngữ tự nhiên.

LSTM hoạt động thông qua ba cổng chính, mỗi cổng kiểm soát dòng chảy thông tin trong ô bộ nhớ tại mỗi bước thời gian. Cổng quên quyết định mức độ thông tin từ trạng thái ô trước đó sẽ được giữ lại, được tính bằng công thức:

$$f_t = \sigma(U_f x_t + W_f h_{t\text{-}1} + b_f) \quad (6)$$

Trong đó, f_t là cổng quên, xt là đầu vào tại thời điểm t, $h_{t\text{-}1}$ là trạng thái ẩn của bước thời gian trước, và các tham số U_f, W_f, b_f được học trong quá trình huấn luyện. Cổng đầu vào xác định mức độ thông tin mới sẽ được đưa vào bộ nhớ và được tính như sau:

$$i_t = \sigma(U_i x_t + W_i h_{t\text{-}1} + b_i) \quad (7)$$

Cổng đầu ra kiểm soát mức độ thông tin từ trạng thái ô sẽ được đưa ra lớp tiếp theo, được tính bằng:

$$o_t = \sigma(U_o x_t + W_o h_{t\text{-}1} + b_o) \quad (8)$$

Ngoài ra, trạng thái ô ứng viên, $c\tilde{}t$, được tính bằng công thức:

$$c\tilde{}_t = \tanh(U_c x_t + W_c h_{t\text{-}1} + b_c) \quad (9)$$

Trong bài toán dự báo tiêu thụ điện năng này, mô hình LSTM đóng vai trò như một thành phần trung tâm giúp mô hình hiểu và khai thác các mối quan hệ phụ thuộc theo thời gian – điều mà các mô hình học máy truyền thống như XGBoost không thực hiện được. Khi lượng điện tiêu thụ tại một thời điểm cụ thể thường phụ thuộc vào các thời điểm trước đó, khả năng ghi nhớ của LSTM trở nên vô cùng quan trọng. Thay vì chỉ dựa vào các đặc trưng tĩnh như ngày, giờ hay thời tiết, LSTM cho phép

mô hình học được chu kỳ tiêu thụ theo ngày, theo tuần, hay thậm chí cả những xu hướng dài hạn. Ví dụ, mô hình có thể nhận biết rằng vào buổi tối lượng tiêu thụ điện thường cao hơn, hay vào những ngày cuối tuần có xu hướng giảm nhẹ.

Cấu trúc đặc biệt của LSTM – bao gồm các cổng ghi nhớ (memory gates) và các trạng thái ẩn (hidden states) – giúp nó vừa ghi nhớ thông tin quan trọng từ quá khứ xa, vừa phản ứng kịp thời với những thay đổi gần nhất trong chuỗi dữ liệu. Trong trường hợp này, đầu vào của LSTM không chỉ là giá trị tiêu thụ điện trước đó, mà còn được kết hợp với đầu ra của mô hình XGBoost – vốn được huấn luyện từ chuỗi các đặc trưng thống kê. Nhờ vậy, mô hình LSTM có thể dự báo chính xác hơn bằng cách tổng hợp cả kiến thức học được từ XGBoost và khả năng nắm bắt thông tin chuỗi của chính nó.

Kết quả là một hệ thống lai, nơi XGBoost cung cấp những đặc trưng tổng quát từ chuỗi đầu vào, và LSTM tiếp nhận những đặc trưng đó để xử lý sâu hơn theo chiều thời gian. Như vậy, LSTM đóng vai trò là lớp xử lý sau cùng, chịu trách nhiệm chính trong việc đưa ra dự báo dựa trên toàn bộ ngữ cảnh đã được học, từ quá khứ đến hiện tại.

3.4 Mô hình XGBoost LSTM

Trong bài toán dự báo chuỗi thời gian, việc chỉ sử dụng một mô hình đơn lẻ thường khó có thể khai thác đầy đủ cả đặc trưng thống kê và tính phụ thuộc theo thời gian của dữ liệu. Chính vì vậy, mô hình lai kết hợp giữa XGBoost và LSTM đã được đề xuất như một giải pháp hiệu quả để tận dụng thế mạnh của cả hai phương pháp.

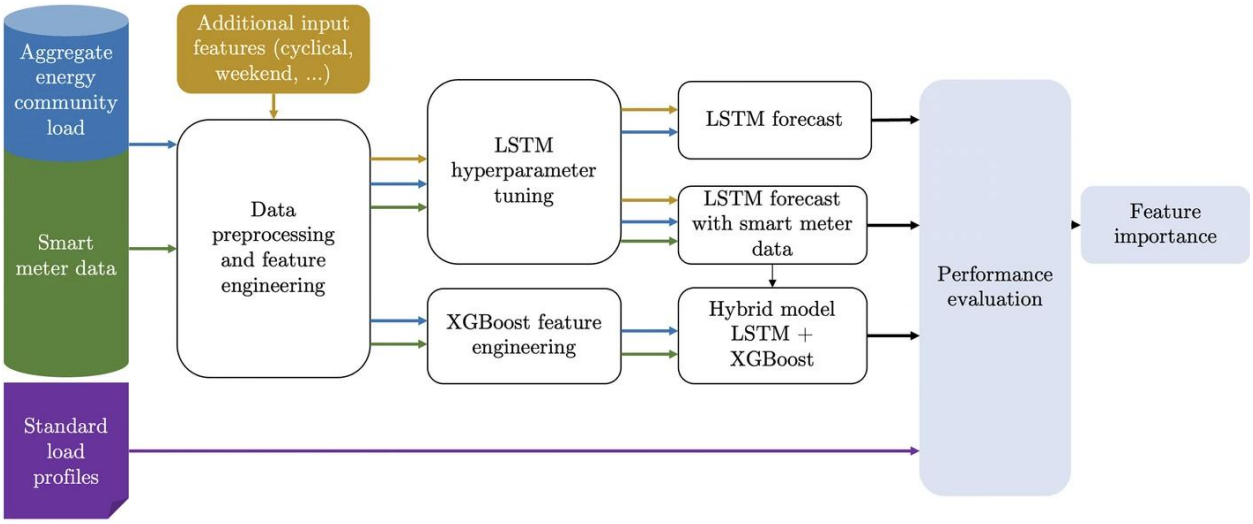
Ý tưởng cốt lõi của mô hình này là sử dụng XGBoost như một bộ trích xuất đặc trưng phi tuyến mạnh mẽ từ chuỗi thời gian đã được mã hóa thành các chuỗi trượt, sau đó truyền đầu ra của XGBoost làm đầu vào mở rộng cho mô hình LSTM. Cụ thể, chuỗi dữ liệu ban đầu (ví dụ: tiêu thụ điện năng theo giờ) được chuẩn hóa bằng phương pháp Min-Max để đưa về cùng một thang đo [0, 1], đảm bảo sự ổn định trong quá trình huấn luyện. Sau đó, dữ liệu được chuyển thành dạng chuỗi trượt (sliding window), với mỗi chuỗi con có độ dài bằng số bước trễ (lag) đã định.

Tiếp theo, mô hình XGBoost được huấn luyện để dự đoán giá trị kế tiếp của chuỗi, dựa trên toàn bộ cửa sổ trễ. Đầu ra của XGBoost – về mặt bản chất là một giá trị dự đoán – được coi như một đặc trưng phi tuyến bổ sung. Đặc trưng này sau đó được kết hợp cùng giá trị gần nhất trong chuỗi đầu vào, tạo thành một vector gồm hai chiều: một chiều là thông tin chuỗi gần nhất, chiều còn lại là đầu ra từ XGBoost. Kết

quả là một tensor đầu vào có dạng (samples,1,2), phù hợp với yêu cầu kiến trúc của LSTM trong PyTorch.

LSTM lúc này đóng vai trò như một bộ dự báo chuỗi, tiếp nhận đầu vào hai chiều để học được cả xu hướng ngắn hạn (từ điểm gần nhất) và thông tin phi tuyến mang tính tổng quát (từ XGBoost). Với cấu trúc có khả năng ghi nhớ dài hạn và loại bỏ nhiễu thời gian ngắn, LSTM giúp tinh chỉnh đầu ra cuối cùng, từ đó cải thiện đáng kể độ chính xác của dự báo.

Cách tiếp cận lai này đặc biệt phù hợp trong các bài toán chuỗi thời gian có tính phi tuyến cao và phụ thuộc thời gian mạnh như dự báo năng lượng, nhu cầu, hay chỉ số thị trường. Trong quá trình thực nghiệm, mô hình XGBoost-LSTM đã cho thấy hiệu quả vượt trội so với từng mô hình đơn lẻ, cả về độ chính xác dự báo lẫn khả năng bắt được chu kỳ biến động dài hạn.



Hình 1. Sơ đồ tổng quan mô hình XGBoost-LSTM

Sơ đồ tổng quan mô hình XGBoost-LSTM thể hiện một hệ thống lai nhằm cải thiện độ chính xác dự báo phụ tải năng lượng bằng cách kết hợp các nguồn dữ liệu khác nhau và tận dụng ưu điểm của hai mô hình mạnh: LSTM và XGBoost.

Quá trình bắt đầu từ ba nguồn dữ liệu đầu vào: (1) dữ liệu phụ tải tổng cộng của cộng đồng, (2) dữ liệu từ công tơ thông minh (smart meter data), và (3) các hồ sơ tải tiêu chuẩn (standard load profiles). Ngoài ra, mô hình còn sử dụng các đặc trưng bổ sung như thông tin chu kỳ (ngày trong tuần, cuối tuần, ngày lễ, v.v.) để nâng cao tính biểu đạt của dữ liệu.

Toàn bộ dữ liệu đầu vào được xử lý thông qua bước tiền xử lý và trích xuất đặc trưng (feature engineering), nơi dữ liệu được làm sạch, chuẩn hóa và xây dựng lại theo định dạng chuỗi thời gian có thể sử dụng được bởi mô hình học sâu.

Song song, có hai nhánh xử lý:

- Nhánh đầu tiên tập trung vào tinh chỉnh siêu tham số của LSTM, huấn luyện mô hình để dự báo phụ tải dựa trên chuỗi dữ liệu đã xử lý, cả với và không có dữ liệu từ công tơ thông minh.
- Nhánh thứ hai sử dụng kỹ thuật trích xuất đặc trưng bằng XGBoost để học các mối quan hệ phi tuyến sâu sắc giữa các đặc trưng đầu vào và giá trị đầu ra dự báo.

Kết quả từ hai nhánh này được tích hợp trong mô hình lai (hybrid model) – kết hợp đầu ra từ LSTM với đặc trưng hoặc đầu ra từ XGBoost, tạo nên một hệ thống mạnh mẽ vừa học được động lực thời gian (qua LSTM) vừa khai thác được cấu trúc dữ liệu phi tuyến (qua XGBoost).

Cuối cùng, mô hình được đánh giá thông qua các tiêu chí hiệu suất dự báo. Đồng thời, tính quan trọng của từng đặc trưng đầu vào cũng được phân tích, giúp hiểu rõ yếu tố nào ảnh hưởng lớn đến dự báo, từ đó phục vụ cho việc tối ưu hoặc ra quyết định trong thực tế.

4 Thực nghiệm và đánh giá

4.1 Dataset

Bộ dữ liệu ECL chứa thông tin về mức tiêu thụ điện năng của 321 khu vực khác nhau được lấy từ bài báo [3]. Gồm 26304 dòng, 322 trường. Dữ liệu được ghi lại theo từng giờ. Thời gian bao phủ khoảng 3 năm, từ 01/01/2012 đến 31/12/2015.

Các trường dữ liệu:

- date: Trường này là dấu thời gian (timestamp), cho biết thời điểm ghi nhận mức tiêu thụ điện năng. Định dạng là "năm-tháng-ngày giờ: phút: giây". Ví dụ: "2012-01-01 00:00:00".
- Các trường MT_000, MT_001, MT_002, ..., MT_320: Mỗi trường này đại diện cho một khu vực khác nhau và chứa giá trị số, thể hiện mức tiêu thụ điện năng của khu vực đó vào giờ tương ứng trong cột date.

Ý nghĩa của dữ liệu: Bộ dữ liệu này cung cấp cái nhìn chi tiết về sự biến động của tiêu thụ điện năng theo thời gian và không gian. Nó cho phép chúng ta:

- Phân tích xu hướng tiêu thụ điện của từng khu vực.

- So sánh mức tiêu thụ điện giữa các khu vực khác nhau.
- Tìm ra các mẫu tiêu thụ điện theo giờ, ngày, mùa.
- Xây dựng mô hình dự báo nhu cầu điện trong tương lai.

Thông tin này rất quan trọng cho các công ty điện lực, nhà quản lý năng lượng và các nhà nghiên cứu để đưa ra các quyết định thông minh về quản lý nguồn cung, phân phối điện và quy hoạch phát triển năng lượng.

4.2 Tiền xử lý

Dữ liệu đầu vào được lấy từ tập tin ECL.csv, với cột dữ liệu chính là MT_000, đại diện cho lượng tiêu thụ điện năng theo thời gian. Để đảm bảo mô hình học hiệu quả và hội tụ nhanh, dữ liệu được chuẩn hóa về khoảng [0,1] bằng phương pháp Min-Max Scaling, sử dụng công cụ MinMaxScaler từ thư viện sklearn.

Sau đó, để mô hình học được tính chất thời gian của chuỗi, dữ liệu được xử lý theo kỹ thuật cửa sổ trượt (sliding window**)**. Cụ thể, mỗi đầu vào gồm 120 bước thời gian trước đó (tức lag = 120), và đầu ra là giá trị tiêu thụ tại bước tiếp theo. Quá trình này tạo thành các cặp mẫu đầu vào - đầu ra (X, y) phù hợp cho cả mô hình XGBoost và LSTM.

Tuy nhiên, vì XGBoost yêu cầu dữ liệu đầu vào ở dạng bảng 2 chiều, các chuỗi đầu vào X ban đầu có kích thước (samples, 120, 1) được chuyển đổi thành dạng phẳng (samples, 128). Mô hình XGBoost được huấn luyện trên tập đặc trưng này để học dự đoán điểm tiếp theo trong chuỗi.

Tiếp theo, để khai thác khả năng học phi tuyến mạnh của LSTM và đồng thời tận dụng thông tin học được từ XGBoost, đầu ra dự đoán từ XGBoost được sử dụng như một đặc trưng bổ sung. Cụ thể, đặc trưng XGBoost được ghép với điểm cuối cùng trong chuỗi đầu vào, tạo thành đầu vào mới có hai đặc trưng cho mỗi bước thời gian:

- Đặc trưng thứ nhất là giá trị tiêu thụ gần nhất,
- Đặc trưng thứ hai là dự đoán từ mô hình XGBoost.

Kết quả cuối cùng là một tensor đầu vào có dạng (samples, 1,2), phù hợp với yêu cầu của kiến trúc LSTM trong PyTorch.

4.3 Cài đặt thực nghiệm

Trong phần thực nghiệm, tôi sử dụng tập dữ liệu Electricity Consumption Load (ECL) để dự báo mức tiêu thụ điện năng tại trạm MT_001. Dữ liệu được chuẩn hóa bằng Min-Max Scaler để đưa về khoảng [0, 1] nhằm phục vụ cho cả mô hình XGBoost và LSTM.

Để khai thác thông tin theo chuỗi thời gian, tôi xây dựng dữ liệu đầu vào bằng cách tạo các chuỗi con với độ trễ (lag) là 120 bước thời gian. Các chuỗi này được sử dụng để huấn luyện mô hình XGBoost, đóng vai trò như một bộ trích xuất đặc trưng đầu tiên. Sau khi huấn luyện, đầu ra của XGBoost (dự báo ngắn hạn) được kết hợp với điểm dữ liệu cuối cùng trong chuỗi lag để tạo thành đầu vào có 2 chiều đặc trưng cho mô hình LSTM.

Mô hình LSTM được thiết kế đơn giản với một lớp LSTM có 64 đơn vị ẩn và một lớp tuyến tính đầu ra. Toàn bộ mô hình được huấn luyện trong 20 epoch bằng bộ tối ưu Adam với learning rate là 0.001 và hàm mất mát là MSELoss.

Sau khi huấn luyện, mô hình được sử dụng để dự báo toàn bộ tập dữ liệu huấn luyện, giúp đánh giá khả năng khớp dữ liệu (fitting) ban đầu. Các kết quả được đưa về thang đo gốc để trực quan hóa và đánh giá bằng các chỉ số như MAE, RMSE, MAPE và R^2.

Ngoài ra, tôi thực hiện dự báo chuỗi 5000 bước tiếp theo dựa trên dãy đầu vào cuối cùng trong tập dữ liệu. Ở mỗi bước, mô hình sử dụng đầu ra của bước trước đó như đầu vào tiếp theo theo cơ chế trượt. Kết quả được hiển thị bằng biểu đồ cũng như được làm mượt bằng trung bình trượt để quan sát xu hướng.

Cuối cùng, để đánh giá tính ổn định theo từng độ dài dự báo khác nhau, chúng tôi tiến hành đo hiệu suất của mô hình tại các horizon khác nhau (12, 24, 36 và 48 bước). Với mỗi horizon, mô hình thực hiện 100 lần dự báo từ các điểm khác nhau trong chuỗi thời gian. Các chỉ số đánh giá trung bình được tính toán để phản ánh hiệu năng tổng quát của hệ thống trong các điều kiện dự báo ngắn và trung hạn.

4.4 Đánh giá

Để đánh giá hiệu quả của mô hình XGBoost-LSTM trong dự báo, tôi đã sử dụng ba chỉ số phổ biến: MAE, RMSE và MAPE và một số chỉ số khác. Đây là những thước đo giúp phản ánh mức độ sai lệch giữa giá trị mô hình dự đoán và giá trị thực tế, từ đó đưa ra cái nhìn toàn diện về độ chính xác và tính ổn định của mô hình.

Chỉ số đầu tiên là **MAE – *Mean Absolute Error*,** hay sai số tuyệt đối trung bình. Đây là một chỉ số đơn giản nhưng rất trực quan. MAE được tính bằng cách lấy trung bình tổng các sai số tuyệt đối giữa dự đoán và thực tế, theo công thức:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

Chỉ số này cho thấy trung bình mỗi dự đoán lệch bao nhiêu so với thực tế, bất kể là lệch nhiều hay ít. MAE vì thế thường được dùng để đánh giá mức độ sai số một cách ổn định và dễ hiểu.

Chỉ số thứ hai là **RMSE** – *Root Mean Squared Error*. Nếu như MAE nhìn nhận sai số một cách bình đẳng, thì RMSE lại đặc biệt nhấn mạnh những sai số lớn. Bằng cách bình phương sai số trước khi tính trung bình rồi lấy căn bậc hai, RMSE khiến cho các giá trị dự đoán sai nghiêm trọng sẽ ảnh hưởng nhiều hơn đến kết quả đánh giá:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{11}$$

Điều này giúp RMSE trở thành một công cụ nhạy hơn trong việc phát hiện những điểm bất thường hoặc dự đoán lệch quá xa so với thực tế.

Cuối cùng là **MAPE** – *Mean Absolute Percentage Error*, hay sai số phần trăm tuyệt đối trung bình. Đây là chỉ số rất dễ diễn giải vì nó biểu thị sai số dưới dạng phần trăm, giúp người đọc dễ dàng hiểu được mức độ lệch của mô hình so với thực tế:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{12}$$

Nhờ ba chỉ số này, tôi đã đánh giá hiệu suất mô hình một cách toàn diện – từ mức độ sai số trung bình, độ nhạy với những sai lệch lớn, cho đến khả năng truyền tải thông tin dự đoán một cách trực quan.

4.5 Kết quả và so sánh

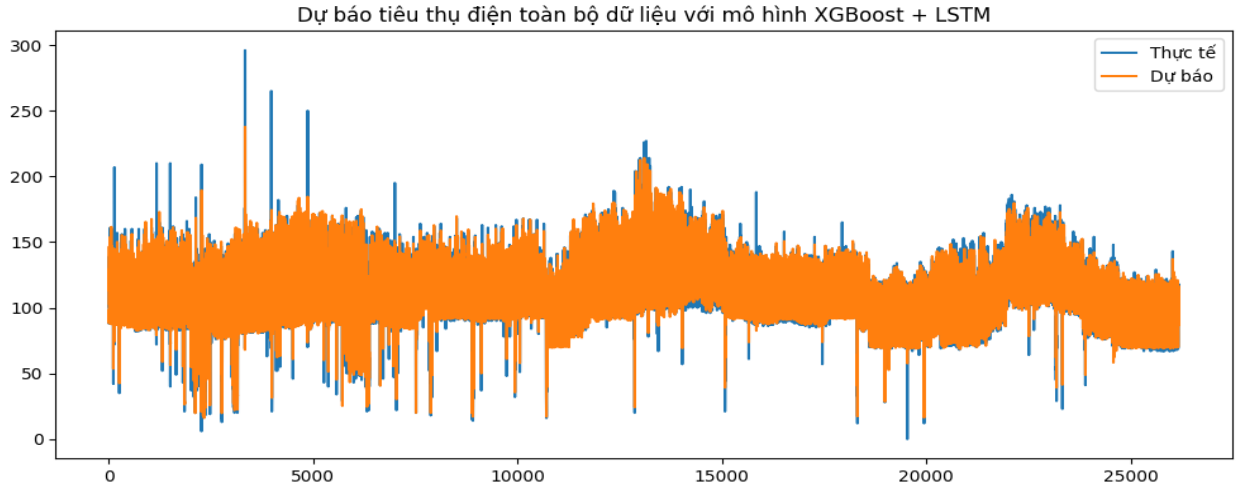
4.5.1 Kết quả của mô hình XGBoost-LSTM

Kết quả thực nghiệm của mô hình XGBoost-LSTM trên tập dữ liệu ECL với lag = 120 được thể hiện qua bảng sau:

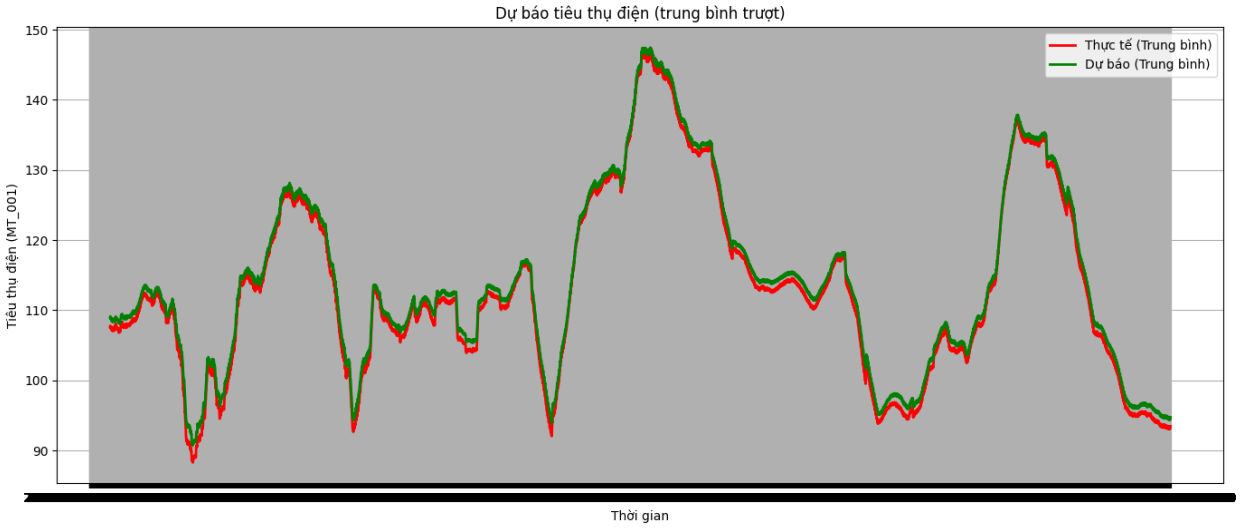
Model	MAE	MSE	RMSE	MAPE	R^2
XGBoost-LSTM	4.5057	48.9799	6.9986	4.46%	0.9251

Bảng 1. Chỉ số đánh giá hiệu suất mô hình XGBoost-LSTM của khu vực MT_001.

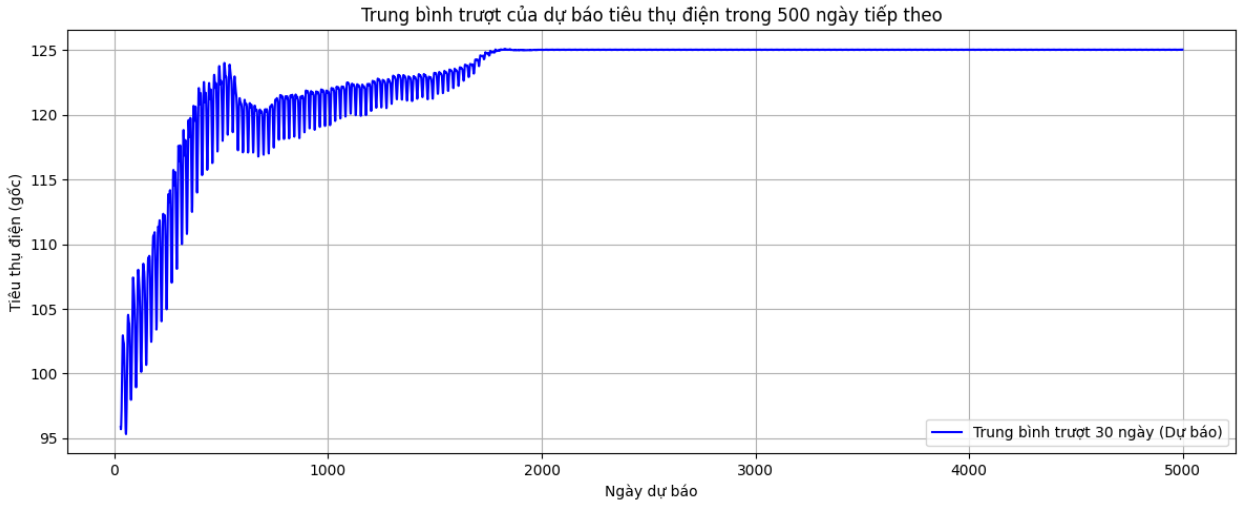
Bảng 1 cho thấy mô hình đạt được độ chính xác khá cao trong bài toán dự báo tiêu thụ điện. Cụ thể, sai số tuyệt đối trung bình (MAE) là 4.5057, nghĩa là trung bình mỗi điểm dự báo chỉ lệch khoảng 4.5 đơn vị so với giá trị thực tế. Sai số bình phương trung bình (MSE) là 48.9799, thể hiện tổng sai số đã được bình phương nhằm nhấn mạnh các sai lệch lớn. Từ đó, sai số căn bình phương trung bình (RMSE) được tính là 6.9986, giúp thể hiện mức độ sai số trung bình theo cùng đơn vị với dữ liệu gốc. Về mặt phần trăm sai số, chỉ số MAPE đạt mức 4.46%, cho thấy mô hình dự báo lệch trung bình khoảng hơn một phần tư so với giá trị thực tế – một mức sai số tương đối chấp nhận được trong các bài toán thời gian thực có nhiều biến động. Đặc biệt, hệ số xác định R^2 đạt 0.9251, chứng tỏ mô hình giải thích được tới hơn 92% phương sai trong dữ liệu thực tế, điều này cho thấy mô hình có độ phù hợp cao và hoàn toàn có thể ứng dụng cho các bài toán dự báo tương tự trong thực tiễn.



Hình 2. Biểu đồ kết quả dự báo tiêu thụ điện trên toàn bộ khu vực MT_001



Hình 3. Biểu đồ trung bình trượt mức độ tiêu thụ điện của khu vực MT_001



Hình 4. Biểu đồ dự báo lượng điện sẽ được tiêu thụ trong 500 ngày tiếp theo

Qua những kết quả đã thực nghiệm ở trên mô hình XGBoost-LSTM cho thấy hiệu quả ở mức trung bình trong việc dự báo tiêu thụ điện. Nó có khả năng nắm bắt các xu hướng và chu kỳ chung của dữ liệu. Tuy nhiên, mô hình gặp một số hạn chế trong việc tái tạo chính xác các biến động ngắn hạn và dự đoán các thay đổi dài hạn. Nhìn chung, hiệu suất của mô hình duy trì sự ổn định trên các khoảng thời gian dự báo ngắn.

4.5.2 So sánh và bàn luận

Để đánh giá hiệu quả tổng thể của mô hình đề xuất, tôi thực hiện so sánh với một số mô hình hiện đại khác trong lĩnh vực dự báo chuỗi thời gian. Các mô hình được lựa

chọn đều có kết quả đáng chú ý trên các benchmark phổ biến và được công bố gần đây, bao gồm:

- iTransformer (2024): Kiến trúc attention cải tiến chuyên biệt cho dữ liệu chuỗi thời gian dài.
- RLinear (2023): Mô hình tuyến tính đơn giản nhưng hiệu quả, đặc biệt tối ưu cho các tập dữ liệu lớn.
- DLinear (2023): Mô hình phân tách tuyến tính theo xu hướng và thành phần nhiễu, nổi bật trong việc giảm độ phức tạp.
- TimesNet (2023): Kiến trúc convolutional-hybrid với khả năng học biểu diễn chuỗi phức tạp đa tần số.
- XGBoost-LSTM (Ours): Mô hình lai do chúng tôi đề xuất, kết hợp khả năng trích chọn đặc trưng phi tuyến của XGBoost với khả năng học động lực chuỗi của LSTM.

Các mô hình được so sánh trên cùng một tập dữ liệu, sử dụng các horizon dự báo lần lượt là 12, 24, 36 và 48 bước thời gian. Mỗi mô hình được huấn luyện và đánh giá lặp lại nhiều lần để đảm bảo tính ổn định của kết quả. Các chỉ số đánh giá bao gồm MAE (Mean Absolute Error), giúp phản ánh toàn diện cả sai số tuyệt đối lẫn tương đối, cũng như mức độ giải thích phương sai dữ liệu.

Model	XGBoost-LSTM(Ours)		iTransformer (2024)		Rlinear (2023)		Dlinear (2023)		TimesNet (2023)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	335.4	12.41	0.178	0.27	0.219	0.298	0.212	0.3	0.268	0.365

Bảng 2. So sánh kết quả dự báo trên tập dữ liệu ECL giữa mô hình đề xuất và các mô hình hiện đại với Horizon S = {12, 24, 36, 48}

Trong bảng 2 khi so sánh mô hình XGBoost-LSTM (Ours) với các mô hình khác trong bảng, chúng ta có thể thấy rằng mô hình của bạn có MSE và MAE cao hơn đáng kể, cho thấy mức độ sai lệch trong các dự báo của nó. Cụ thể, MSE của XGBoost-LSTM là 335.4 và MAE là 12.41, cho thấy rằng mô hình có độ phân tán và sai số dự báo cao hơn so với các mô hình khác. Điều này có thể phản ánh sự phức tạp của việc kết hợp hai kỹ thuật khác nhau (XGBoost và LSTM) để dự báo chuỗi thời gian, dẫn đến sai số lớn hơn.

Trong khi đó, mô hình iTransformer (2024) có MSE cực kỳ thấp (0.178) và MAE chỉ là 0.270, cho thấy khả năng dự báo chính xác rất cao. Đây là một mô hình vượt trội, với khả năng nắm bắt các phụ thuộc thời gian phức tạp một cách hiệu quả hơn so với XGBoost-LSTM.

Các mô hình Rlinear (2023) và Dlinear (2023) cũng thể hiện kết quả khá tốt với MSE lần lượt là 0.219 và 0.212, và MAE lần lượt là 0.298 và 0.300, tuy có sai số thấp hơn so với XGBoost-LSTM, nhưng vẫn chưa thể đạt được mức độ chính xác như iTransformer. Tuy nhiên, chúng vẫn có hiệu suất tốt hơn XGBoost-LSTM trong việc dự báo chính xác các giá trị trung bình.

TimesNet (2023) có MSE là 0.268 và MAE là 0.365, vẫn thấp hơn XGBoost-LSTM, nhưng không thể so sánh với iTransformer về độ chính xác. Tuy nhiên, với những chỉ số này, TimesNet có thể được coi là một lựa chọn cạnh tranh trong một số trường hợp.

Tổng thể, mặc dù XGBoost-LSTM có MAE chấp nhận được, nhưng MSE của nó cao cho thấy rằng mô hình này vẫn còn khoảng cách lớn so với các mô hình khác, đặc biệt là iTransformer, trong việc đạt được độ chính xác cao trong dự báo chuỗi thời gian. Việc so sánh với các mô hình như Rlinear, Dlinear, và TimesNet cho thấy rằng mặc dù XGBoost-LSTM có thể không phải là mô hình chính xác nhất, nhưng vẫn có tiềm năng trong việc dự báo, đặc biệt trong các bài toán yêu cầu sự kết hợp giữa các kỹ thuật học máy truyền thống và học sâu.

5 Kết luận

Trong nghiên cứu này, tôi đã xây dựng và triển khai mô hình LSTM nhằm dự báo mức tiêu thụ điện năng theo giờ dựa trên dữ liệu lịch sử. Kết quả thực nghiệm cho thấy mô hình có khả năng học được các đặc trưng thời gian dài hạn của chuỗi dữ liệu, nhờ vào cơ chế ghi nhớ của mạng LSTM. Thử nghiệm với độ trễ (lag) khác nhau cho thấy việc lựa chọn giá trị lag phù hợp - ví dụ như 120 giờ - có ảnh hưởng rõ rệt đến hiệu suất dự báo, do phản ánh được chu kỳ tiêu thụ ngắn hạn trong ngày và trong tuần.

Mặc dù mô hình đạt kết quả khá tốt trên tập dữ liệu hiện tại, vẫn còn một số hạn chế như chưa khai thác thêm các đặc trưng ngoài chuỗi (ví dụ: thời tiết, ngày lễ, giá điện...). Trong tương lai, tôi dự định mở rộng mô hình bằng cách tích hợp các đặc trưng này, áp dụng thêm các mô hình tiên tiến như Transformer hoặc hybrid LSTM-

CNN, và triển khai mô hình trong môi trường thực tế để kiểm nghiệm khả năng dự báo theo thời gian thực.

6 Một số công trình liên quan

Trong công trình [3], tác giả đề cập đến sự phát triển của các mô hình dự báo chuỗi thời gian, đặc biệt là việc áp dụng mô hình Transformer, vốn ban đầu được thiết kế cho xử lý ngôn ngữ tự nhiên, vào việc nắm bắt các phụ thuộc thời gian phức tạp. Tác giả chỉ ra rằng hiệu quả của các mô hình Transformer trong dự báo chuỗi thời gian hiện đang bị nghi ngờ, nhất là khi so sánh với các mô hình tuyến tính đơn giản hơn. Điều này đã dẫn đến sự phát triển của các biến thể Transformer, từ những thay đổi về cơ chế attention như trong Autoformer và Informer, đến các phương pháp toàn diện hơn kết hợp các đặc tính chuỗi thời gian như tính dừng và patching. Cụ thể, công trình này giới thiệu mô hình iTransformer, trong đó tác giả đảo ngược việc áp dụng các thành phần của Transformer hiện có, với hy vọng đây là cách tiếp cận hiệu quả hơn đối với kiến trúc Transformer trong bài toán dự báo chuỗi thời gian.

Trong công trình [4], tác giả đề cập đến sự phát triển và tầm quan trọng của việc nghiên cứu tải sạc xe điện, đặc biệt là trong bối cảnh cần phải lập kế hoạch sạc xe điện hiệu quả trong hệ thống điện lưới. Tác giả nhấn mạnh rằng việc dự báo tải sạc chính xác là yếu tố quan trọng để tối ưu hóa việc điều phối điện năng, giúp cân bằng sự khác biệt giữa các mức đỉnh và đáy của tải điện và cải thiện hiệu suất sử dụng nguồn cung cấp điện. Tác giả cũng chỉ ra rằng các phương pháp truyền thống sử dụng mô hình xác suất như mô hình trung bình xác suất và mô phỏng Monte Carlo có những hạn chế, đặc biệt là sự phụ thuộc vào độ chính xác của các thống kê xác suất và khó khăn trong việc thích ứng với các thay đổi trong quy tắc sạc giữa các khu vực. Do đó, tác giả chuyển hướng sang các phương pháp dựa trên dữ liệu, sử dụng các kỹ thuật như XGBoost và LSTM để nhận diện các mẫu thời gian phức tạp, nhằm cải thiện độ chính xác và khả năng ứng dụng trong dự báo tải sạc xe điện.

Tài liệu tham khảo

1. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735–1780.
2. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794)
3. Liu, Yong, et al. "itransformer: Inverted transformers are effective for time series forecasting." *arXiv preprint arXiv:2310.06625* (2023).
4. Xue, M., Wu, L., Zhang, Q. P., Lu, J. X., Mao, X., & Pan, Y. (2021). Research on load forecasting of charging station based on XGBoost and LSTM model. In *Journal of Physics: Conference Series* (Vol. 1757, No. 1, p. 012145). IOP Publishing.