Phân tích dữ liệu chuỗi thời gian trong bài toán đánh giá và dự đoán khả năng hồi phục của bệnh nhân đột quỵ

Trần Lê Phương Thảo
thao.tlp200604@sis.hust.edu.vn

Lê Văn Bằng
banglv1@viettel.com.vn

Trong bản báo cáo này, tập trung phân tích dữ liệu chuỗi thời gian yếu tố vật lý của các bài tập hồi phục ngoài chỉ số y học để xây dựng mô hình dự đoán và đánh giá khả năng hồi phục trong tương lai. Đầu tiên, phân tích dữ liệu để trích xuất ra những đặc trưng thống kê như tính dừng của chuỗi, độ trễ tối ưu và môi quan hệ tương quan giữa các biến để xây dựng mô hình hồi quy vec-tơ (Vector Autoregressive Model). Ngoài ra, cách tiếp cận với mô hình dự báo khác như sử dụng mô hình học máy LightGBM và mô hình học sâu Long short-term memory (LSTM). Kết quả cho thấy mô hình LSTM kết hợp với phương pháp đơn phổ (Singular Spectrum Analysis) phân tách thành phần chuỗi thời gian cho ra độ chính xác của mô hình dự báo cao nhất. Đối với bài toán ước lượng khả năng hồi phục của bệnh nhân sử dụng kết quả dự đoán trong tương lai bằng mô hình phân loại như Binary Classification và Forecasting Ensemble dựa trên ngưỡng xác định trước.

Keyword. *Vector Autoregressive Model (VAR), LightGBM, Long short-term memory (LSTM), Singular Spectrum Analysis (SSA)*

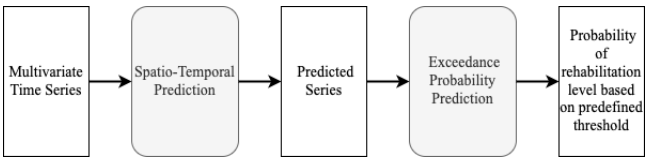
Viettel Digital Talent 2022 . . .

1. INTRODUCTION

Cải thiện khả năng đi lại là mục tiêu chung của những người sau đột quỵ và là trọng tâm chính của việc hồi phục chức năng. Các bước đi bộ và tốc độ đo được trong bài kiểm tra tiêu chuẩn 6 Minutes Walking Test (6MWT) đại diện cho sự tham gia bên ngoài môi trường lâm sàng, do đó cung cấp cái nhìn sâu sắc về hoạt động thể chất và mức độ phục hồi của bệnh nhân. Mục đích của nghiên cứu hiện tại là: (1) dự báo số bước và tốc độ đi bộ dựa vào chuỗi thời gian trong quá khứ; (2) ước lượng khả năng hồi phục của bệnh nhân trong khoảng thời gian được xác định trong tương lai.

Bài toán dự đoán số bước và tốc độ đi bộ dựa trên chuỗi thời gian trong quá khứ được coi là bài toán **Spatio-Temporal Prediction**. Vì thế, việc phân tích Granger Causality Test là cần thiết để xác định mối tương quan giữa các biến dữ liệu để xác định những đặc điểm quan trọng của mô hình. Và các tiếp cận tiếp theo là xây dựng mô hình dự báo số bước đi và tốc độ đạt hiệu quả và độ chính xác cao. Hiện nay, theo các lý thuêts khác nhau về xây dựng mô hình dự báo chuỗi thời gian, các mô hình dự báo có thể chia thành 3 loại: mô hình time series, mô hình Machine Learning và mô hình Deep Learning. Các mô hình chuỗi thời gian sẽ được ứng dụng trong bài toán như **Auto-Regressive Itergrated Moving Average** (ARIMA) và **Vector Auto-Regressive** (VAR). Mô hình học máy đã dần được áp dụng và không ngừng đưa ra một số mô hình mới để dự đoán xu hướng tương lai, chẳng hạn như **LightGBM**. Học sâu là một công cụ hiện đại để trích xuất và dự đoán dữ liệu một cách tự động. Nó có khả năng thích ứng mạnh mẽ và khả năng tự học mà không cần phải xác định tương quan giữa các biến dữ liệu và các mô hình toán học cụ thể. Tuy nhiên, một số mô hình truyền thống như **Multiple Layer Perceptron** (MLP) dễ gặp phải vấn đề overfitting và phụ thuộc vào mối quan hệ tương quan giữa các biến trong dữ liệu, nhưng mạng **Recurrent Neural Network** (RNN) có thể giải quyết những vấn đề này. Nhưng vấn đề của RNN gặp phải như là vanishing gradient và có thể không hội tụ về tối ưu. Vì vậy, trong bài toán này sẽ sử đụng mạng **Long short-term memory** (LSTM) thay thế cho RNN để khắc phục vấn đề này.

Bài toán ước lượng khả năng hồi phục của bệnh nhân trong khoảng thời gian được xác định trong tương lai được coi là bài toán dự báo xác suất vượt quá **Exceedance Probability Prediction**. Dự báo xác suất vượt quá là vấn đề ước tính xác suất mà một chuỗi thời gian sẽ vượt quá ngưỡng xác định trước trong một khoảng thời gian tương lai được xác định trước. Các nhân viên y tế sẽ dựa trên các ước tính về việc liệu số bước và tốc độ đi bộ có vượt qua ngưỡng mà người khoẻ mạnh đạt được để điều chỉnh pháp đồ điều trị cho bệnh nhân. Đầu ra theo xác suất là mong muốn vì nó mang nhiều thông tin hơn để hỗ trợ việc ra quyết định. Dự báo "khả năng hồi phục trong 3 tháng tới của bệnh nhân A là 10%" sẽ mang nhiều thông tin hơn là chỉ nói đơn giản "trong 3 tháng tới bệnh nhân A chưa thể hồi phục hoàn toàn". Trong bản báo cáo, sẽ đề xuất 2 cách tiếp cận chính được sử dụng để giải quyết các vấn đề về dự báo xác suất vượt quá là **Binary Classification** và **Forecasting Ensemble**.



Hình 1. Mô tả về bài toán

Bản báo cáo này được chia thành 5 phần. Phần 1 là giới thiệu cơ sở nghiên cứu; phần 2 là phân tích dữ liệu time series; phần 3 là giới thiệu các phương pháp luận cho xây dựng mô hình dự báo; phần 4 là so sánh các kết quả thực nghiệm của từng mô hình; phần 5 là kết luận và thảo luận cho hướng phát triển của nghiên cứu.

2. DATA ANALYSIS

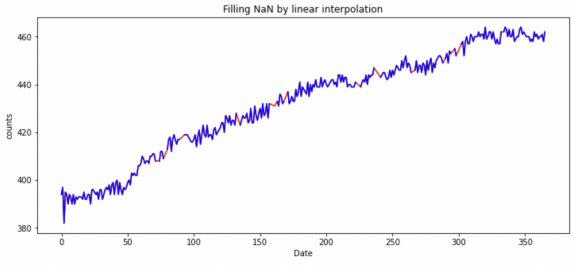
2.1. *Dataset*

Tập dữ liệu được giả lập từ STRGDE dataset (Stroke Initiative for Gait Data Evaluation). Dữ liệu thu được từ 2 nhóm đối tượng: post-stroke patients và healthy people từ 5 trường đại học tại Mỹ ghi lại kết quả của bài kiểm tra người từng bị đột quy và người khoẻ mạnh với bài tập đi bộ trong 4 phút cho ra các đặc trưng: vận tốc trung bình, thời gian được timestamp của từng bước, độ dài của sải bước,...
Tuy nhiên, dữ liệu của tập dữ liệu gốc chỉ được ghi lại ở một thời điểm, không có sẵn chuỗi thời gian, nên dữ liệu được sử dụng để làm cơ sở cho khởi tạo dữ liệu thời điểm đầu của chuỗi thời gian hợp lý có thể. Cách giả lập tập dữ liệu RSPDC (Recovery of Stroke Patient based on Daily-step Counts Data)cho bài toán bằng cách tạo time series cho 1 đối tượng, dùng mô hình TimeGAN (Time-series Generative Adversarial Network) để tạo cho các đối tượng còn lại.
Tóm tắt về tập dữ liệu RSPDC dùng trong bài toán. Dữ liệu gồm 31 đối tượng (N=31) là người đã từng bị đột quy và thực hiện bài tập 6 Minute Walking Test (6MWT) cho quá trình phục hồi được quan sát trong thời gian 1 năm được thu thập từ ngày 1/1/2020 đến ngày 31/12/2020 bởi các đặc trưng: số bước đi (`counts`) + vận tốc (`speed`)

2.2. *Data Extraction*

Gap Filling Techniques

Bộ dữ liêu ghi lại có số lượng NaN trên 1 đối tượng trong khoảng (13-92) thời điểm, nguyên nhân gây thiếu dữ liệu có thể là nguyên nhân chủ quan của đối tượng khi không thực hiện bài tập phục hồi. Và khoảng cách trống của dữ liệu trong khoảng (1-7) thời điểm. Phương pháp đề xuất để giải quyết vấn đề điền khoảng trống trong chuỗi dự liệu là một số phương pháp nội suy như linear interpolate, spline interpolate. Việc khoảng cách trống dữ liệu quá lớn sẽ ảnh hưởng đến kết quả của phép nội suy.
Ngoài ra, phương pháp lấp đầy khoảng trống của SSA có thể được sử dụng. Đối với một chuỗi thời gian đơn biến, quy trình lấp đầy khoảng trống SSA sử dụng các tương quan thời gian để điền vào các điểm còn thiếu. Đối với tập dữ liệu đa biến, việc lấp đầy khoảng trống bằng M-SSA tận dụng đuọc cả tương quan không gian và thời gian (spatio-temporal).



Hình 2. Phương pháp xử lý NaN bằng cách linear interpolation. Với những điểm màu đỏ là giá trị NaN được được điền bằng nội suy tuyến tính.

Kiểm định nghiệm đơn vị

Chuỗi dừng có ý nghĩa quan trọng trong lý thuyết đồng liên kết. Vì thế, trong quá trình ước lượng các tham số hoặc kiểm định giả thiết của các mô hình, nếu không kiểm định thuộc tính dừng thì các kỹ thuật phân tích quan hệ nhân quả giữa các biến hay áp dụng mô hình hồi quy dự đoán sẽ không còn chính xác và hợp lý.
Đối với dữ liệu chuỗi hồi phục y tế, các chuỗi này thường không dừng, vì vậy để tạo ra chuỗi dừng cần phải lấy sai phân. Để xem xét chuỗi dừng hay không sử dụng kiểm định là ADF (Augmented-Dickey Fuller) như ví dụ ở bảng 1.
Kết quả kiểm định cho thấy cả hai chuỗi không dừng ở mức

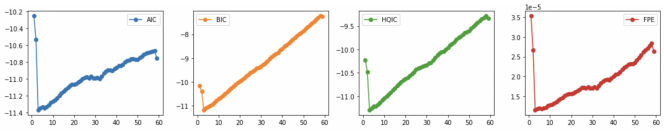
	Chuỗi ban đầu		Chuỗi sai phân bậc 1	
	counts	speed	counts	speed
p-value	0.9956	0.8343	0.0000	0.0000
critical value-1%	0.9956	0.8343	0.0000	0.0000
critical value-5%	-2.8696	-2.8695	-2.8695	-2.8695
critical value-10%	0.9956	0.8343	0.0000	0.0000

Bảng 1. **Kiểm định ADF về tính dừng của chuỗi thời gian của 2 biến counts và speed.**

ý nghĩa 1%, 5% và 10%. Mặt khác xác suất để bác bỏ H0 với độ tin cậy 95% là $p - value_{counts} = 0.9956 > 0.05$ và $p - value_{speed} = 0.8343 > 0.05$. Như vậy không đủ cơ sở để bác bỏ H0 cả 2 chuỗi ban đầu là chuỗi dừng. Vậy cần xem xét chuỗi sai phân bậc 1 có tính dừng hay không. Kết quả cho thấy cả hai chuỗi sai phân bậc 1 đều dừng ở mức ý nghĩa 1%, 5% và 10%. Vì vậy chuỗi dừng bậc 1 của cả 2 biến sẽ được sử dụng để thay thế cho chuỗi ban đầu.

Kiểm định độ trễ

Kiểm tra độ trễ tối đa nhằm xác định độ trễ của chuỗi dữ liệu khi kiểm định mô hình nhân quả và độ trễ của các biến trong mô hình hồi quy xác định mức tác động. Trong các tiêu chuẩn kiểm định độ trễ bao gồm , BIC, HQIC và FPE, sử dụng tiêu chuẩn AIC để kiểm tra và xác định độ trễ tối ưu khi AIC nhận giá trị nhỏ nhất.



Hình 3. Biểu đồ thể hiện giá trị của các tiêu chuẩn AIC, BIC, HQIC và FPE thay đổi trong khoảng đỗ trễ được lựa chọn là 60. thu được đỗ trễ tối ưu là p=3.

Kiểm định nhân quả Granger

Granger Causality Test được sử dụng để kiểm chứng chiều hướng tác động của các cặp biến để xem xét trong các viến, biến nào là biến nguyên nhân, biến nào là biến kết quả. Biến Y_t được cho là có mối quan hệ nhân quả với biến Xt nếu dữ liệu quá khứ của Xt có thể dùng để dự báo cho biến Y_t.
Kiểm định nhân quả Granger nhiều biến được thực hiện bằng cách điều chỉnh mô hình VAR với L là độ trễ tối ưu như biểu diễn sau:

$$X(t) = \sum_{\tau=1}^{L} A_\tau X(t-\tau) + \epsilon(t)$$

ở đây $\epsilon(t)$ là white Gaussian random vector, A_τ là ma trận cho mọi giá trị τ. Một chuỗi thời gian X_i được gọi là nguyên nhân Granger của một chuỗi thời gian khác X_j nếu ít nhất một trong các phần tử $A_\tau(i,j)$ với mọi $\tau = 1, ..., L$ lớn hơn đáng kể so với 0 (so sánh về giá trị tuyệt đối).
Xác định các điều kiện cần trước khi thực hiện kiểm định nhân quả Granger:

- Các biến cần kiểm định nhân quả là các chuỗi dừng hoặc không có tương quan giả
- Chiều hướng của mối quan hệ nhân quả phụ thuộc vào lựa chọn độ trễ.
- Các phần dư (residual component) không có hiện tượng tự tương quan (nếu có thì cần chuyển sang phân tích sai phân của chuỗi)

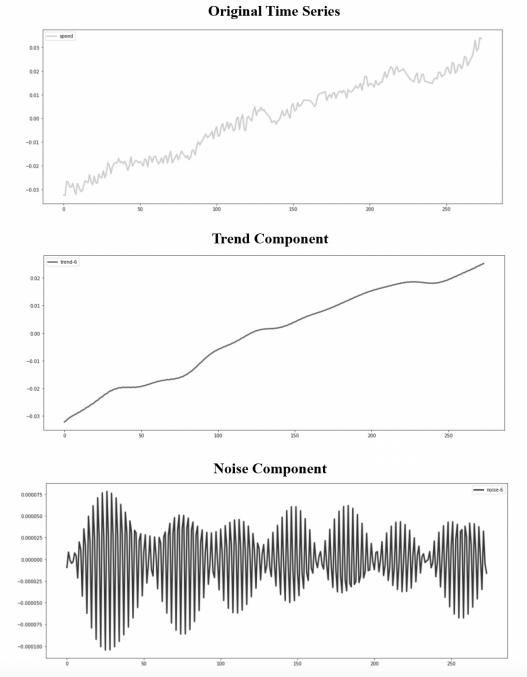
	counts-X	speed-X
counts-Y	1.000000	**0.000000**
speed-Y	**0.000904**	1.000000

Bảng 2. **Granger Causation Matrix.** Trong trường hợp này, cả 2 giá trị $p-value$ đều < 0.05 nên có thể kết luận cả 2 biến đều là nguyên nhân Granger của nhau (speed $\Longleftrightarrow$ counts)

2.3. *Time Series Decomposition*

Singular Spectrum Analysis

Phân tích đơn phổ (SSA) được sử dụng như một kỹ thuật không có mô hình để nó có thể được áp dụng cho các chuỗi thời gian tùy ý bao gồm cả chuỗi thời gian không có tính dừng. Mục đích cơ bản của SSA là phân tích chuỗi thời gian thành tổng các thành phần có thể khai thác được xu hướng, phát hiệu chu kỳ, điều chỉnh theo mùa, làm mịn và giảm thành phần nhiễu.



Hình 4. Phân tách chuỗi thời gian. Chuỗi thời gian ban đầu của biến tốc độ được phân tách thành 2 thành phần là xu hướng (trend) và phần nhiễu (noise).

3. METHODOLOGY

3.1. *Vector Auto-Regressive Model (VAR)*

Một hạn chế của các mô hình hồi quy như ARIMA là chúng áp đặt mối quan hệ một chiều trên 1 biến, nghĩa là biến dự báo bị ảnh hưởng bởi những biến dự báo trước thời điểm đó. Tuy nhiên, có trường hợp có phép tất cả các biến ảnh hưởng lẫn nhau. Qua kiểm định nhân quả Granger, có thể phân tích được mối quan hệ hai chiều giữa 2 biến, cụ thể như những thay đổi trong số bước đi được dự báo dựa trên những thay đổi trong vận tốc của đối tượng, và ngược lại.

Mô hình VAR là sự tổng quát hóa của mô hình ARIMA cho đơn biến để dự báo một vector chuỗi thời gian. Biểu diễn phương trình cảu mô hình VAR như sau:

Trên thực tế, một yếu điểm của mô hình VAR là bản thân mô hình không được xây dựng trên lý thuyết của lĩnh vực nào được áp đặt cấu trúc vào các phương trình. Tất cả các biến được giả định sẽ ảnh hưởng đến mọi biến khác trong cả một hệ thống nhất, nên việc giải thích các hệ số ước tính trở nên khó khăn. Mặc dù vậy, mô hình VAR có tính hữu ích trong hoàn cảnh: 1) dự báo tập hợp các biến liên quan mà không cần giải thích rõ ràng; 2) làm cơ sở cho kiểm định nhân quả Granger xem một biến có liên quan đến việc dự báo biến khác hay không.

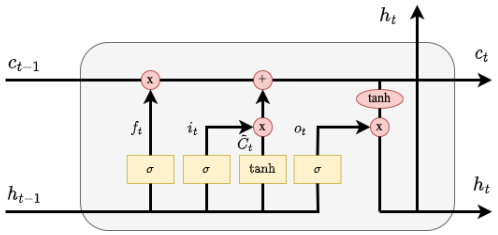
3.2. *LightGBM*

Để dự báo với mô hình học máy LightGBM, điều đầu tiên là chuyển dữ liệu time series thành dữ liệu dạng bảng với các đặc trưng được tạo là các giá trị độ trễ tại thời điểm đó $(y_{t-1}, y_{t-2}, y_{t-3}, \dots)$

3.3. *Long short-term memory (LSTM)*

LSTM là một dạng đặc biệt của Recurrent Neural Network (RNN), nó có ưu điểm hơn RNN là khắc phục được vấn đề vanishing gradient. Luồng dữ liệu và cquá trình tính toán của cấu trúc 3 cổng của mạng LSTM được phân tích chi tiết như sau:

Forget Gate Mạng LSTM tính giá trị f từ 0 đến 1 cho $H-1$ và



Hình 5. LSTM Cell.

X, sử dụng giá trị f đề quyết định xem có lưu giữ thông tin của $C-1$ hay không (nếu bằng 0 có nghĩa là loại bỏ hoàn toàn, 1 có nghĩa là lưu hoàn toàn). Giá trị f của forget gate được tính như sau:

$$f_t = (W_f \cdot [h_{t-1}, x_t] + b_f)$$

Input Gate

Hàm kích hoạt sigmoid và tanh được kết hợp để tạo ra một trạng thái cập nhật mới. Hàm điều khiển của cổng vào như sau:

$$i_t = (W_t \cdot [h_{t-1}, x_t] + b_i)$$
$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_c] + b_c)$$

Cellular State

Phương pháp cập nhật cụ thể là kết hợp giá trị cổng quên f với trạng thái ô cũ $C-1$, với C là giá trị ứng viên mới để xác định có bao nhiêu giá trị trạng thái cần được cập nhật. Hàm cập nhật trạng thái từng ô như sau:

$$C_t = f_t \cdot C_{t-1} + i_C \cdot \tilde{C}_t$$

Output Gate

Kết quả đầu ra o được tính bởi cổng đầu ra của LSTM và trạng thái ô C tại thời điểm t được xử lý bởi hàm kích hoạt tanh. Hàm điều khiển của cổng đầu ra như sau:

$$o_t = (W_o \cdot [h_{t-1}, x_t] + b_o)$$
$$h_t = o_t tanh(C_t)$$

Univariate LSTM

Nếu chuỗi thời gian chỉ có 1 biến sẽ được biểu diễn như sau:

$$X_t = \{s_{t-W}, s_{t-W+1}, ..., s_{t-2}, s_{t-1}\}$$

trong đó W là chiều rộng của cửa sổ quan sát.

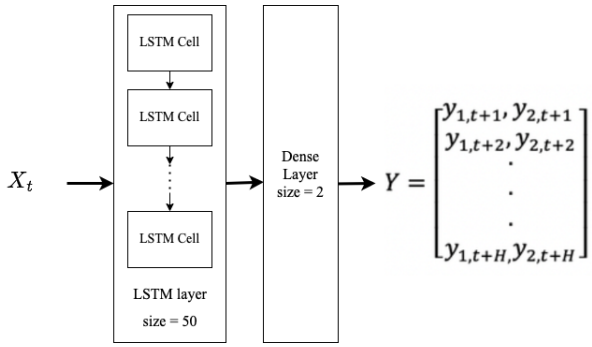
Multivariate LSTM

Nếu số lượng biến là n, biểu diễn chuỗi thời gian đầu vào như sau:

$$S^{[i]} = \{[i]_1^{[i]}, [i]_2^{[i]}, ..., [i]_T^{[i]}\}, i = 1, 2, ..., N$$

trong đó i đại diện cho biến dữ liệu, dạng ma trận đầu vào tại thời điểm t được biểu diễn như sau:

$$X_t = \begin{bmatrix} s_{t-W}^{[1]} & s_{t-W+1}^{[1]} & \cdots & s_{t-1}^{[1]} \\ s_{t-W}^{[2]} & s_{t-W+1}^{[2]} & \cdots & s_{t-1}^{[2]} \\ \vdots & \vdots & \ddots & \vdots \\ s_{t-W}^{[N]} & s_{t-W+1}^{[N]} & \cdots & s_{t-1}^{[N]} \end{bmatrix}$$

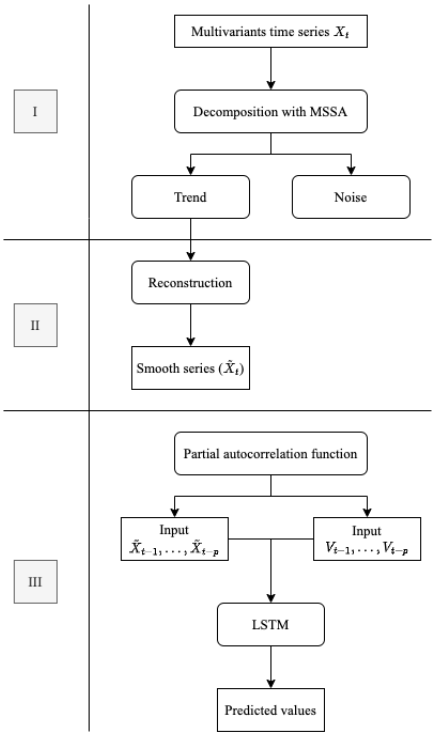


Hình 6. Sơ đồ của mô hình dự báo LSTM cho nhiều biến

3.4. *Combine LSTM with SSA*

Với mục đích để cải thiện hiệu quả của độ chính xác mô hình và khai thác được yếu tố thành phần xu hướng, tiến hành kết hợp tiền xử lý dữ liệu để phân tách chuỗi thời gian thành các yếu tố thành phần và tái cấu trúc thành chuỗi thời gian theo thành phần bằng phương pháp đơn phổ (Singular Spectrum Analysis).

Mô hình kết hợp SSA-LSTM bao gồm 3 stages và quá trình mô hình được miêu tả như sơ đồ hình . Stage 1 là phân tách theo phương pháp đơn phổ (SSA) từ chuỗi dữ liệu thời gian ban đầu thành 2 yếu tố thành phần là xu hướng (trend) và phần nhiễu (noise). Stage 2 là tái cấu tạo chuỗi thời gian sau khi loại bổ phần nhiễu (noise). Stage 3 là sử dụng mô hình dự đoán LSTM, dựa vào hàm tự tương quan một phần PACF (Partial autocorrelation function) để xác định đỗ trễ.



Hình 7. Sơ đồ của mô hình kết hợp SSA và LSTM

3.5. *Binary Classification*

Vượt quá ngưỡng liên quan đến các sự kiện nhị phân. Do đó, cách tự nhiên nhất để lập mô hình xác suất vượt ngưỡng là đóng khung vấn đề dưới dạng phân loại nhị phân. Các biến dùng để huấn luyện đại diện cho các quan sát trong quá khứ, cùng với dữ liệu của của các biến khác. Biến mục tiêu là giá trị nhị phân, cho biết liệu sự kiện có xảy ra hay không . Sau đó, việc ước tính xác suất vượt quá bằng cách sử dụng mô hình phân loại xác suất như Logistic Regression.

3.6. *Forecast Ensemble*

Một cách tiếp cận khác với cách phân loại là sử dụng Forecast Ensemble. Về cơ bản, sẽ huấn luận một số mô hình dự báo để dự đoán các giá trị tương lai của chuỗi thời gian. Sau đó, ước tính xác suất vượt ngưỡng bằng cách tính toán tỷ lệ của giá trị dự báo từ các mô hình dự báo so với giá trị vượt ngưỡng. Để tạo mô hình Forecast Ensemble, sử dụng Random Forest để huấn luận một số Decision Trees trên tập huấn luyện.

4. EXPERIMENTS

4.1. *Data preprocessing*

Chuỗi dữ liệu theo từng biến sau khi Gap Filling có thống kê xác suất theo như bảng . Đối với bài toán dự báo theo chuỗi thời gian và ước lượng khả năng hồi phục, sẽ chia tập dữ liệu thành 2 tập train và tập test. Tập train có dữ liệu thời gian trong khoảng từ 1/1/2020 đến 30/9/2020 và tập test có dữ liệu chuỗi thời gian trong khaongr từ 1/10/2020 đến 31/12/2020.

4.2. *Perfomance Measure*

Spatio-Temporal Prediction

Root mean square error (RMSE), Mean absolute error (MAE) và Mean absolute percentage error (MAPE) được sử dụng như chỉ số chính xác trong dự đoán để đánh giá mức độ hiệu quả của dự đoán trên tập test. Những giá trị của các chỉ số trên càng nhỏ thì độ chính xác của mô hình dự đoán càng cao.

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

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

trong đó y_i là giá trị thực tế, $\hat{y}_i$ là giá trị dự đoán từ mô hình dự báo và N là số lượng thời điểm.

Exceedance Probability Prediction

Đường cong AUC-ROC là phép đo hiệu suất cho các bài toán phân loại ở các ngưỡng thiết lập khác nhau. ROC là một đường cong xác suất và AUC đại diện cho mức đọ hoặc thước đo khả năng phân chia các lớp. AUC càng cao thì mô hình phân loại càng tốt trong việc phân biệt giưã giá trị có vượt ngưỡng hay không vượt ngưỡng. Đường con ROC được xác định bởi giá trị TPR (tương ứng với trục Oy) và giá trị FPR (tương ứng với trục Ox).

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

Điểm Brier là trung bình của bình phương sai số giữa những dự đoán và điểm giá trị thực tương ứng tại thời điểm đó. Sự khác biệt lớn hơn giưã dựu báo xác suất và kết quả sự kiện phản ánh nhiều sai sót hơn trong dự đoán. Do đó, điểm Brier càng tháp cho thấy độ chính xác trong dự đoán càng cao.

$$BrierScore = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

trong đó N là số thời điểm được xem xét, t là chỉ số của thời điểm, f_t là dự báo xác suất cho thời điểm t (f_t nằm trong khoảng từ 0 đến 1), o_t là giá trị thực tại thời điểm t (o_t chỉ nhận giá trị 0 hoặc 1).

4.3. *Results*

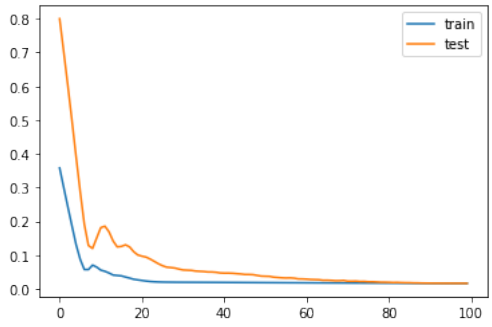
Mô hình dự báo chuỗi thời gian trong tương lai được đánh giá dựa trên độ đo RMSE, MAE và MAPE. Kết quả đánh giá trên tập test được thể hiện cụ thể trong bảng 3.

Kết quả cho thấy mô hình kết hợp SSA-LSTM cho ra các giá trị

Baseline	RMSE		MAE		MAPE	
	counts	speed	counts	speed	counts	speed
VAR	3.800168	0.008569	2.902174	0.007102	0.006333	0.005751
LightGBM	4.376173	0.007773	3.277381	0.006616	0.007158	0.005363
LSTM	2.316863	0.004731	1.928027	0.004129	0.004218	0.003015
SSA-LSTM	**1.442477**	**0.003198**	**1.172822**	**0.003017**	**0.002598**	**0.002541**

Bảng 3. **Kết quả độ đo của mô hình dự báo chuỗi thời gian.**

của độ đo ở tất cả các biến đều có kết quả nhỏ hơn đáng kể so với các mô hình dự báo khác. Chứng tỏ việc phân tách thành phần dữ liệu chuỗi thời gian đóng vai trò rất quan trọng trong việc cải thiện chất lượng mô hình dự báo. Quá trình học của mô hình SSA-LSTM được hội tụ sau khi huấn luyện 100 epochs, thể hiện trong hình 8.

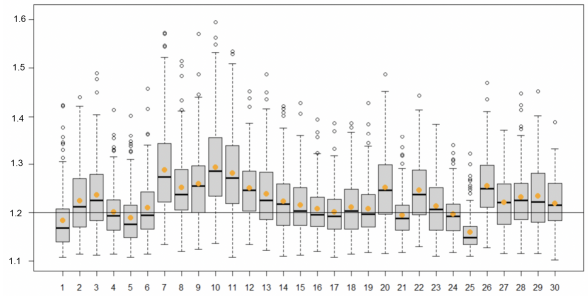


Hình 8. Giá trị hàm loss trong huấn luyện của model SSA-LSTM. Sau 100 epochs giá trị hàm loss của tập train và tập test cùng hội tụ với $loss_{train} = 0.0164$ và $loss_{test} = 0.0162$

Đối với bài toán ước lượng xác suất hồi phục dựa trên ngưỡng xác định trước, trong thực nghiệm tập train được sử dụng cho cả 2 hướng tiếp cận là giá trị dự đoán từ mô hình dự báo kết hợp SSA-LSTM và tập test là giá trị thực của chuỗi dữ liệu biến tốc độ. Trên cơ sở lý thuyết, ngưỡng tốc độ cho thấy mức hồi phục hoàn toàn là $theshold_{speed} = 1.2m/s$. Kết quả so sánh 2 hướng mô hình được thể hiện trong bảng 4.

Baseline	ROC	Brier Score
Binary Classification	**0.918803**	**0.144711**
Forecast Ensemble	0.254273	0.964696

Bảng 4. Kết quả độ đo của mô hình ước lượng xác suất hồi phục dựa trên ngưỡng xác định trước.



Hình 9. Boxplot chuỗi thời gian dự đoán với mean là giá trị dự đoán và stardard deviation được xác định từ độ đo RMSE. Điểm màu cam là giá trị thực của chuỗi thời gian.

5. Conclusion

Trong nghiên cứu này, đã xây dựng các mô hình dự báo số bước và tốc độ đi bộ của bệnh nhân đột quỵ để làm cơ sở ước lượng xác suất hồi phục của bệnh nhân trong khoảng thời gian tương lai xác định trước dựa trên ngưỡng biết trước. Các mô hình dự báo bao gồm các mô hình chuỗi thời gian cổ điển VAR, mô hình học máy LightGBM, mô hình học sâu LSTM. Ngoài ra, việc sử dụng phương pháp tách thành phần dữ liệu Singular Spectrum Analysis được sử dụng cho tiền xử lý dữ liệu để kết hợp với mạng LSTM cho ra độ chính xác dự đoán cao của mô hình. Hướng tiếp bằng mô hình phân loại Binary Classifition cũng cho ra kết quả dự đoán xác suất vượt ngưỡng của chuỗi thời gian được dự đoán trong tương lai với độ đo ROC cao với 91.88% trên tập test.

Tuy nhiên, đối với cả 2 nhánh của bài toán đều gặp phải một số hạn chế. Ở bài toán dự đoán chuỗi thời gian tương lai, với biến dữ liệu số bước đi bộ việc dự đoán không dự đoán được kết quả nguyên mà thay vì dự đoán số thực và lấy kết quả làm tròn. Còn đối với bài toán ước lượng xác suất hồi phục, mô hình đánh giá đang coi dữ liệu huấn luyện là dữ liệu bất định. Nhưng về mặt thống kê, các dữ liệu được dự báo trong tương lai được sử dụng đánh giá như dữ liệu bất định (uncertainty accessment).

Tài liệu

[1] Rob J Hyndman, George Athanasopoulos. *FORECASTING: PRINCIPLES AND PRACTICE*, https://otexts.com/

[2] Vitor Cerqueira. *An Introduction to Exceedance Probability Forecasting*, https://towardsdatascience.com/an-introduction-to-exceedance-probability-forecasting-4c96c0e7772

[3] Qi Tang, Ruchen Shi, Tongmei Fan, Jingyan Huang. *Prediction of Financial Time Series Based on LSTM Using Wavelet Transform and Singular Spectrum Analysis*, https://www.researchgate.net/figure/Hybrid-SSA-LSTM-model-processing-process_fig5_352268379

[4] *Introduction to Time Series Analysis*, https://phdinds-aim.github.io/time_series_handbook/Preface/Preface.html

[5] Tomonori Masui *Multi-step Time Series Forecasting with ARIMA, LightGBM, and Prophet*, https://towardsdatascience.com/multi-step-time-series-forecasting-with-arima-lightgbm-and-prophet-cc9e3f95dfb0

[6] https://github.com/kieferk/pymssa