Phân tích dữ liệu chuỗi thời gian

trong bài toán đánh giá và dự đoán khả năng hồi phục của bệnh nhân đột quỵ

Viettel Digital Talents

Ngày 24 tháng 6 năm 2022

Presenter: Trần Lê Phương Thảo

Table of contents

01

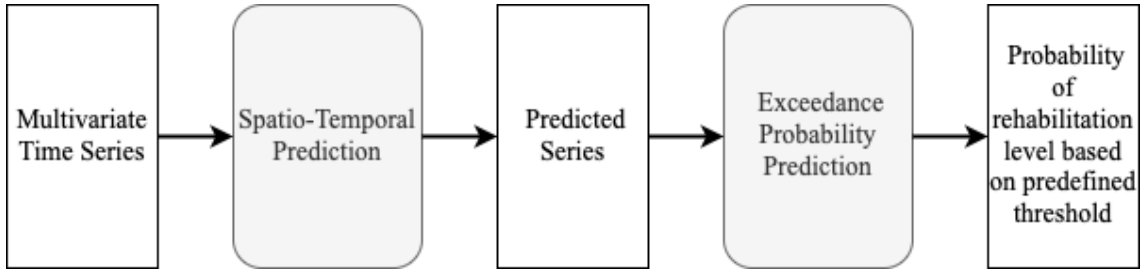
About the problem

About the problem

- Dựa vào **dữ liệu thời gian của các đặc trưng trong môi trường lâm sàng** để dự đoán và đánh giá khả năng hồi phục của bệnh nhân trong tương lai
- Ưu điểm: trợ giúp nhân viên y tế, không cần sự tham gia của các thiết bị y tế
- Hạn chế: chỉ áp dụng được cho một số loại bệnh

- Mục tiêu:

(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

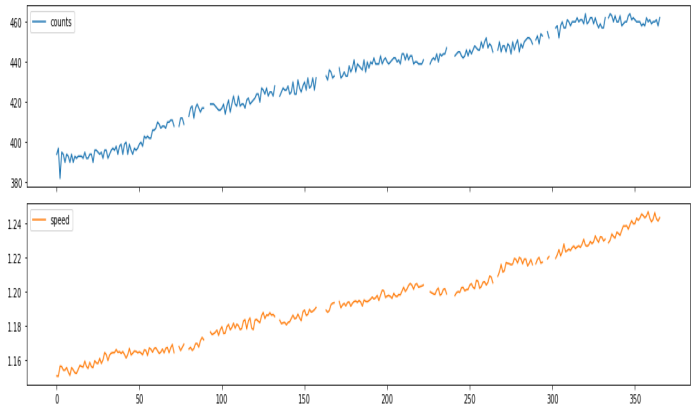


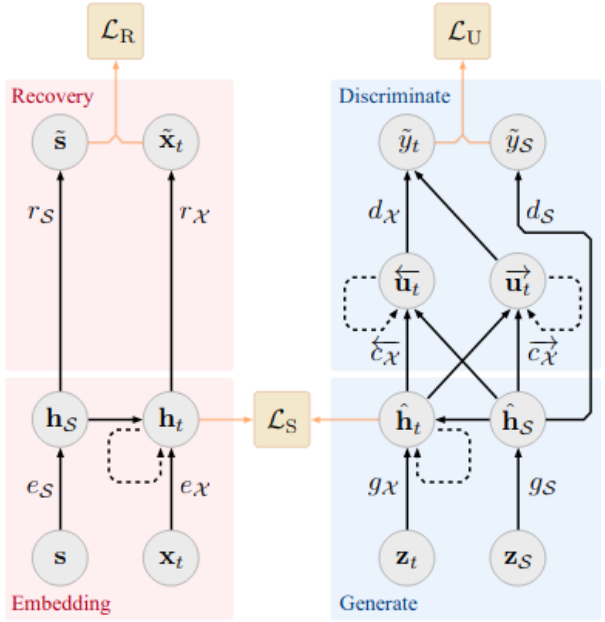
02

Data analysis

Dataset

Tập dữ liệu RSPDC bao gồm chuỗi thời gian về số bước đi và tốc độ đi bộ của 31 đối tượng bệnh nhân đột quỵ.





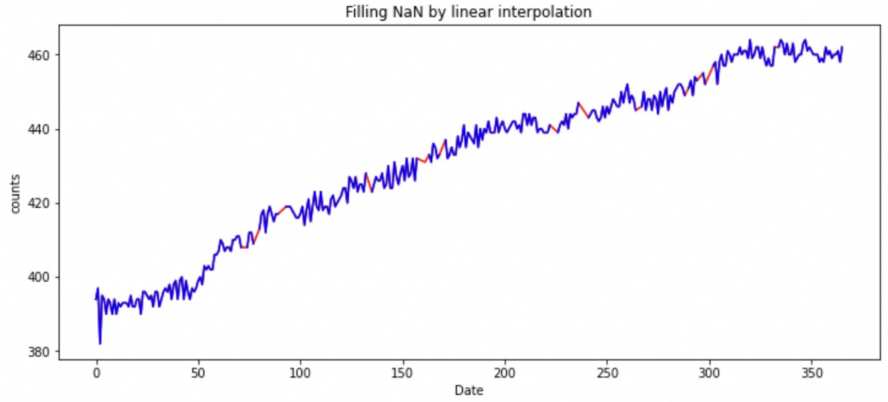
(a) TimeGAN

Source: Synthetic Time-Series Data: A GAN approach | by Fabiana Clemente | Towards Data Science

Data Preprocessing

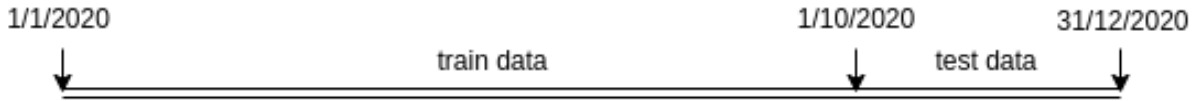
Gap Filling Techniques

Linear Interpolation



Split dataset

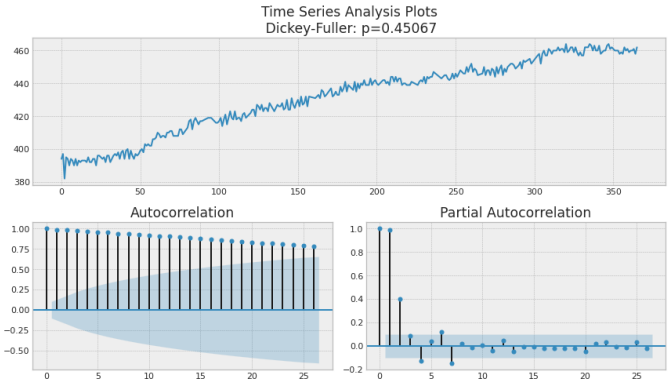
Thực hiện huấn luyện mô hình cho dữ liệu thời gian của 1 đối tượng



Data Extraction

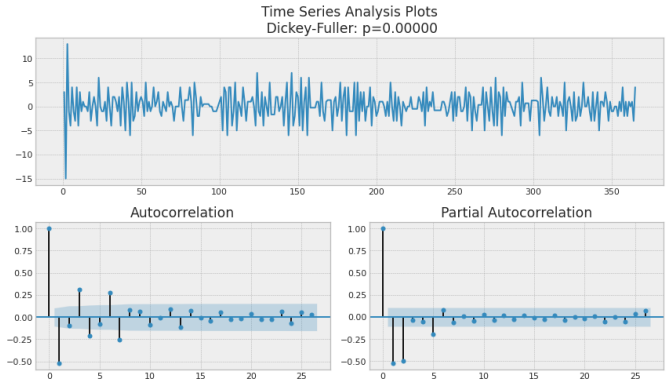
Kiểm định nghiệm đơn vị

Xác định tính dừng của chuỗi bằng kiểm định Augmented Dickey-Fuller (ADF test)



p-value > 0.05

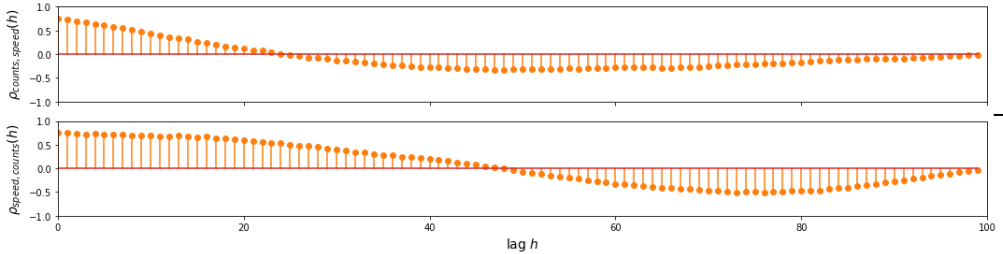
Sai phân bậc 1



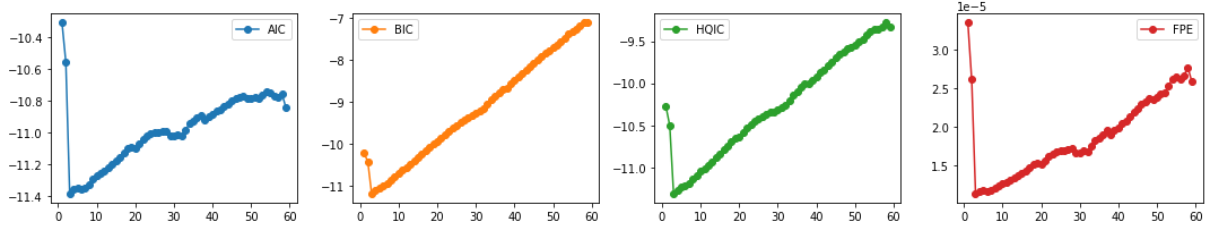
Chuỗi sai phân bậc 1 có tính dừng

Data Extraction

Kiểm định độ trễ (lag)



Mối liên hệ giữa
biến counts - độ trễ của biến speed,
và ngược lại



Độ trễ tối ưu khi chỉ số AIC đạt nhỏ nhất → lag = 3

Data Extraction

Granger Causality Test

Sử dụng để kiểm chứng chiều hướng tác động của các cặp biến để xác định biến nguyên nhân và biến kết quả

Nếu $\sigma^2(X|U) < \sigma^2(X|\overline{U - Y})$,

thì $Y_t \implies X_t$

Dựa vào Granger Causation Matrix, có p-value < 0.05

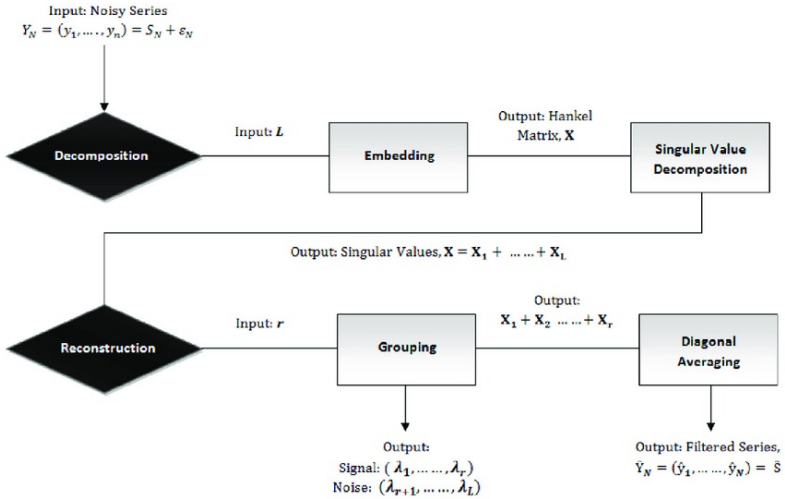
Suy ra $(\text{speed} \iff \text{counts})$

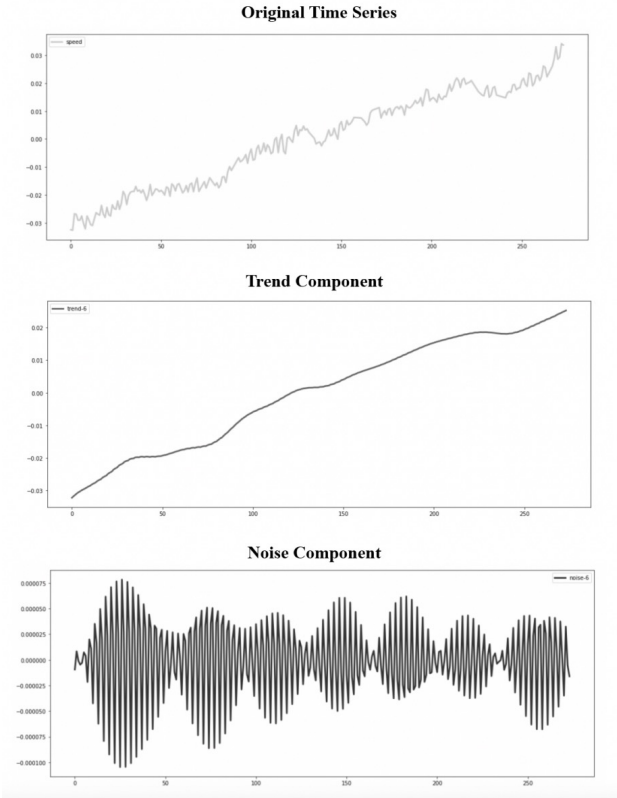
	counts-X	speed_X
counts_Y	1.000000	**0.000000**
counts_Y	**0.000904**	1.000000

Granger Causation Matrix

Time Series Decomposition

Singular Spectrum Analysis (SSA)





Source: The basic singular spectrum analysis (SSA) process [36]. | Download Scientific Diagram (researchgate.net)

03

Methodology

Spatio-Temporal Prediction

Vector Auto-Regressive Model (VAR)

$$\begin{bmatrix} 1 & b_{1,2} \\ b_{2,1} & 1 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} b_{1,0} \\ b_{2,0} \end{bmatrix} + \underbrace{\begin{bmatrix} \varphi_{1,1} & \varphi_{1,2} \\ \varphi_{2,1} & \varphi_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}}_{\text{Auto-Regressive (AR)}} + \underbrace{\begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}}_{\text{Moving Average (MA)}}$$

LightGBM

Supervised Learning Setting

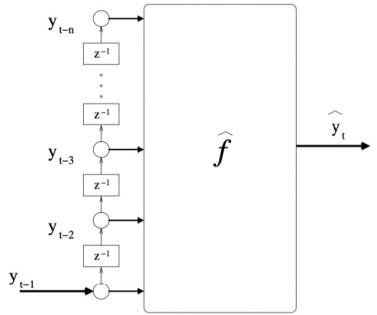
Training Set

$$X = \begin{bmatrix} y_{N-1} & y_{N-2} & \cdots & y_{N-n-1} \\ y_{N-2} & y_{N-3} & \cdots & y_{N-n-2} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & y_{n-1} & \cdots & y_1 \end{bmatrix}$$

Output Vector

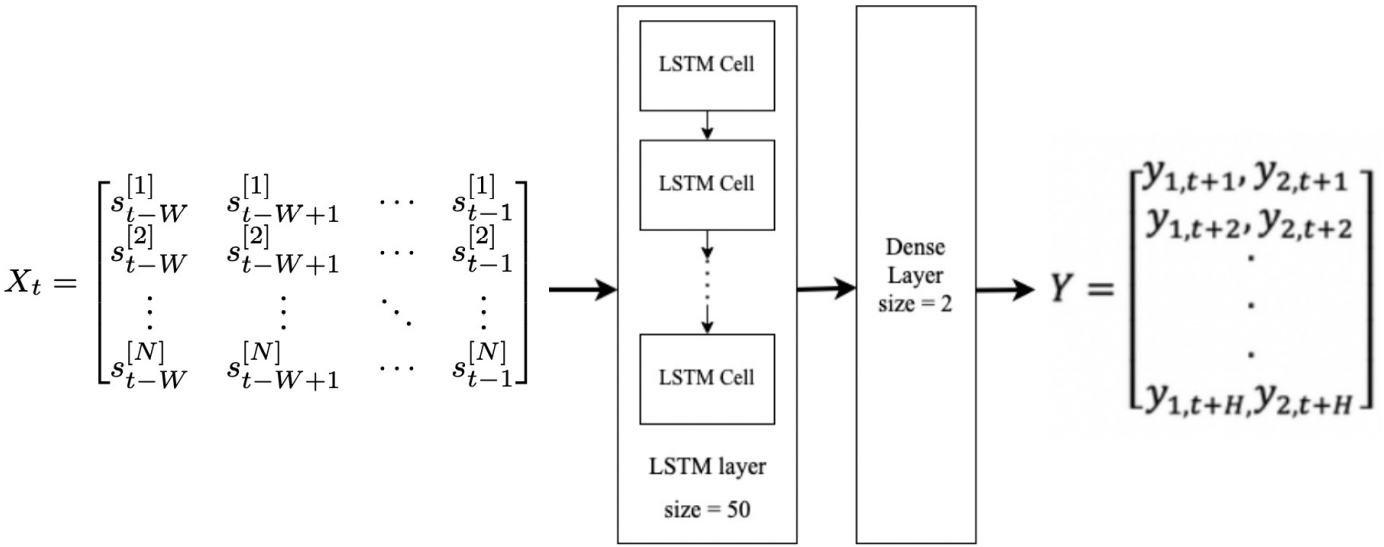
$$Y = \begin{bmatrix} y_N \\ y_{N-1} \\ \vdots \\ y_{n+1} \end{bmatrix}$$

One-step Forecasting



Spatio-Temporal Prediction

Long short-term memory (LSTM)



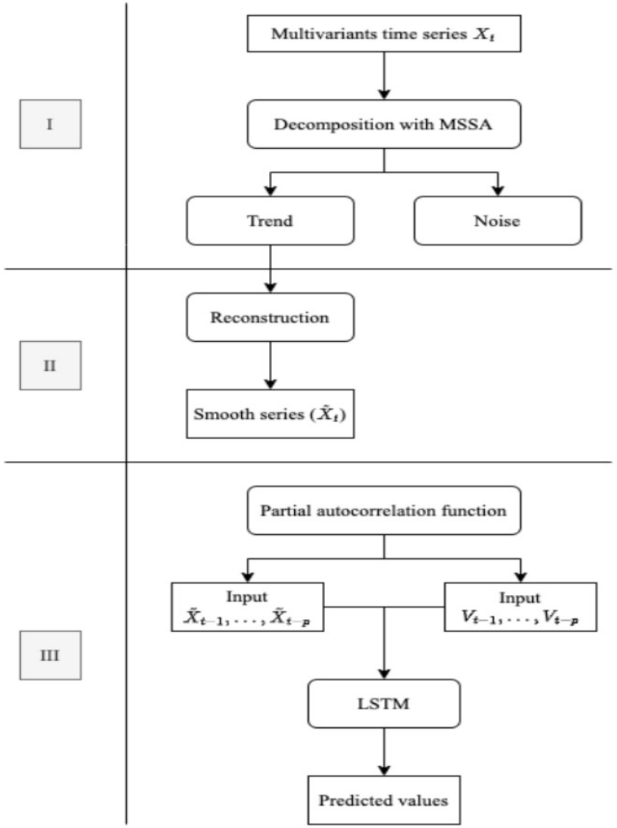
Spatio-Temporal Prediction

Combine LSTM with SSA

Stage 1. 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. Tái cấu tạo chuỗi thời gian sau khi loại bổ phần nhiễu (noise).

Stage 3. Sử dụng mô hình đự đoán LSTM, dựa vào hàm tự tương quan một phần PACF (Partial autocorrelation function) để xác định đỗ trễ.



Exceedance Probability Prediction

Binary Classification

Định nghĩa giá trị của nhãn

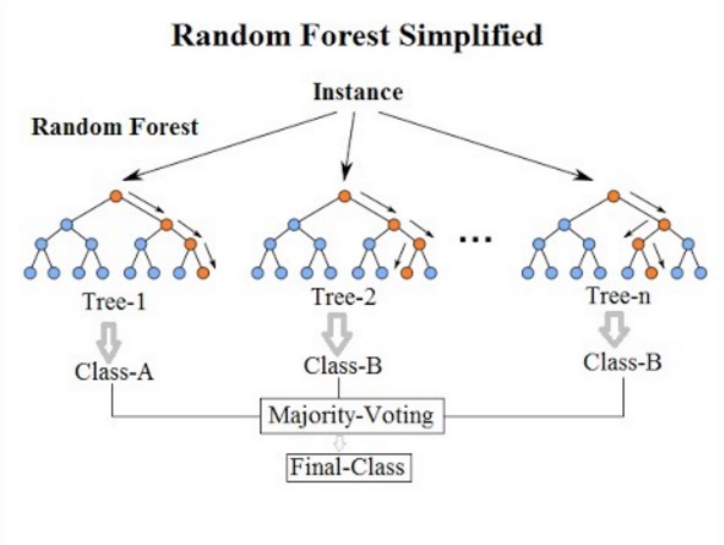
$$X_t = 1 \qquad R_t \geq r$$

$$X_t = 0 \qquad R_t < r \qquad t = 1, \cdots, T$$

Logistic Regression

$$P(R > r) = [1 + \exp(-\boldsymbol{\beta}'\mathbf{z})]^{-1}$$

Forecast Ensemble



Source: Tập tin:Random forest diagram complete.png – Wikipedia tiếng Việt

04

Experiments

Spatio-Temporal Prediction

Perfomance Metrics

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

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

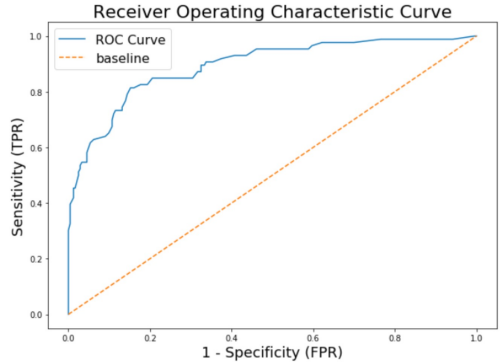
$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Results

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**

Exceedance Probability Prediction

Perfomance Metrics



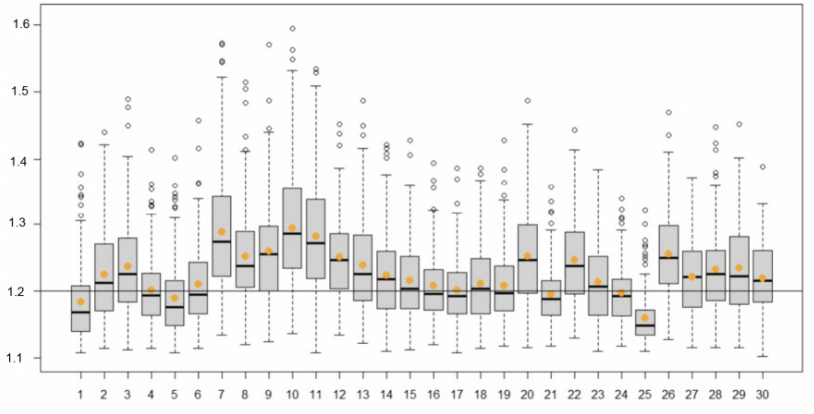
$$TPR = \frac{TP}{TP + FN}$$

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

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

Results

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



05

Conclusion

Conclusion

○ **Bài toán dự đoán số bước đi và tốc độ trong tương lai**

- Phân tích dữ liệu cho các **kiểm định nghiệm đơn vị, độ trễ và nhân quả Granger** là cần thiết cho việc xây dựng mô hình hồi quy VAR
- Time Series Decomposition, đặc biệt là phương pháp **Singular Spectrum Analysis (SSA)** đóng vai trò quan trọng trong việc cải thiện chất lượng dự đoán của mô hình **LSTM**

○ **Bài toán ước lượng xác suất khả năng hồi phục của bệnh nhân**

- Cách tiếp cận **Binary Classification** mang lại độ chính xác dự đoán cao

06

Future Work

Future Work

- **Bài toán dự đoán số bước đi và tốc độ trong tương lai**
  - Mở rộng thêm các biến dữ liệu thời gian như độ dài trung bình bước đi, nhịp tim trong lúc hoạt động
  - Chuỗi thời gian đếm ***counts*** có thể giải quyết vấn đề dự báo nguyên bằng **phương pháp Croston**
  - Sử dụng thêm những lý thuyết và kỹ thuật thống kê trong tiền xử lý như **biến đổi Fourier, biến đổi Wavelet**

- **Bài toán ước lượng xác suất khả năng hồi phục của bệnh nhân**
  - Sử dụng **phương pháp lấy mẫu Boostrap** để tính kết quả xác suất trung bình cho cả khoảng thời gian tương lai xác định trước

Reference

[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

Q&A