BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NIÊN LUẬN NGÀNH
CÔNG NGHỆ THÔNG TIN

Đề tài

PHÂN LOẠI RỐI LOẠN NHỊP TIM BẰNG DỮ LIỆU ĐIỆN TÂM ĐỒ

Sinh viên: ĐẶNG THÀNH TRUNG
Mã số: B1910322
Khóa: K45

Cần Thơ, 12/2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
TRƯỜNG CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN

NIÊN LUẬN NGÀNH
CÔNG NGHỆ THÔNG TIN

Đề tài

PHÂN LOẠI RỐI LOẠN NHỊP TIM BẰNG DỮ LIỆU ĐIỆN TÂM ĐỒ

Người hướng dẫn
TS. PHẠM THẾ PHI

Sinh viên thực hiện
ĐẶNG THÀNH TRUNG
Mã số: B1910322
Khóa: K45

Cần Thơ, 12/2022

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG

DANH MỤC TỪ ĐIỂN – VIẾT TẮT – GIẢI THÍCH THUẬT NGỮ

STT	Viết tắt	Tiếng Anh	Tiếng Việt
1	ECG	Electrocardiogram	Điện tâm đồ
2	MRI	Magnetic Resonance Imaging	Cộng hưởng từ
3	KNN	k-Nearest Neighbors	K láng giềng gần
4	DTW	Dynamic time warping	Sự biến dạng thời gian
5	TSC	Time series classification	Phân loại chuỗi thời gian
6	MLP	Multilayer perceptron	Mạng nơ-ron đa tầng
7	CNN	Convolutional networks	Mạng nơ-ron tích chập
8	RNN	Convolutional networks	Mạng nơ-ron hồi quy
9	XGBoost	Extreme Gradient Boosting	Một dạng của phương pháp tập hợp mô hình
10		Sinus Rhythm	Rối loạn nhịp xoang
11		Atrial Fibrillation	Rung nhĩ
12		Atrioventricular Block	Block nhĩ thất
13		Left Bundle Branch Block	Block nhánh trái
14		Right Bundle Branch Block	Block nhánh phải
15		Premature Atrial Contraction	Ngoại tâm thu nhĩ
16		ST Segment Depression	Đoạn ST chênh lên
17		ST Segment Elevation	Đoạn ST chênh xuống
18		Clustering	Gom cụm
19		Classification	Phân lớp
20		Motif detection	Phát hiện Motif
21		Anomaly detection	Phát hiện chuỗi bất thường

22		Supervised Learning	Học có giám sát
23		Lazy learning	Thuật toán lười học
24		Logistic Regression	Hồi quy logistic
25		Softmax Regression	Hồi quy tối đa
26		Multinomial Logistic Regression	Hồi quy logic đa thức
27		Decision Tree	Cây quyết định
28		Random forest	Rừng ngẫu nhiên
29		Ensemble model	Tập hợp mô hình
30		Bootstrap Aggregating	
31		Gradient Boosting	
32		Cross-validation	
33		Models with hand-engineered features	
34		Models with direct signal input	
35		Epoch	
36		Accuracy	
37		Confusion Matrix	
38		F1 score	

TÓM LƯỢC

Bất thường về nhịp tim là một trong những rối loạn sức khỏe phổ biến trong dân số. Các bất thường này có mức độ nghiêm trọng khác nhau từ khó chịu cho đến mức tử vong. Để phát hiện rối loạn này, đo điện tâm đồ là phương pháp đầu tiên nhất được sử dụng để chẩn đoán. Tuy nhiên, để xác định chính xác tình trạng rối loạn các bác sĩ phải tiến hành phân tích hình ảnh điện tâm đồ; đôi khi có những rối loạn cần theo dõi trong nhiều ngày dẫn đến khả năng quá tải dữ liệu. Với việc áp dụng trí tuệ nhân tạo vào phát hiện bất thường có thể giúp các bác sĩ, chuyên gia giảm được số lượng dữ liệu cần phân tích và tăng được độ chính xác của chẩn đoán. Các mô hình được xây dựng trong phạm vi niên luận có khả năng phân loại các bất thường của nhịp tim bằng dữ liệu điện tâm đồ.

Trong Phần Nội dung, Chương 1 có trình bày ngắn gọn về cơ sở sinh lý của điện tâm đồ và mô tả một số rối loạn. Chương 2 trình bày lý thuyết về học sâu và cách xây dựng một số mô hình học sinh. Chương 3 trình bày về tập dữ liệu được dùng trong việc đào tạo và kiểm tra mô hình.

Các mô hình sử dụng một số thư viện hỗ trợ: NeuroKit2, Pandas, PyTorch, Scikit-Learn, Scipy, XGBoost,…

PHẦN 1. GIỚI THIỆU

I. ĐẶT VẤN ĐỀ

Rối loạn nhịp tim là một bệnh lý phổ biến xảy ra do nhiều nguyên nhân khác nhau và ngày càng có xu hướng trẻ hóa. Rối loạn nhịp tim nếu không được phát hiện và điều trị kịp thời bệnh lý này có thể gây nguy hiểm cho tính mạng người bệnh.

Từ trước đến nay, các bác sĩ chẩn đoán rối loạn nhịp tim bằng nhiều phương pháp khác nhau như dựa vào dấu hiệu, triệu chứng lâm sàng và các thông số cận lâm sàng như: điện tâm đồ (ECG), siêu âm tim, hình ảnh cộng hưởng từ (MRI),… Trong đó điện tâm đồ là một trong những phương pháp được chỉ định đầu tiên trong chẩn đoán. Các bác sĩ dựa vào đồ thị của dòng điện tim được ghi lại trên giấy để phân tích, từ đó tìm ra những bất thường liên quan đến bệnh lý như rối loạn nhịp tim, nhồi máu cơ tim, suy tim cấp, tràn dịch màng ngoài tim,… Tuy nhiên, vấn đề đặt ra là ECG có rất nhiều hình dạng khác nhau đòi hỏi khả năng phân tích và lượng kiến thức rất lớn, ngoài ra thời gian chẩn đoán cũng là một thách thức lớn vì một số bệnh lý tim mạch cần được can thiệp ngay để đảm bảo tính mạng của bệnh nhân.

Ngày nay, dữ liệu về y khoa rất lớn và phức tạp dẫn đến việc bác sĩ gặp khó trong việc tận dụng triệt để các dữ liệu y khoa. Hơn nữa, sự phát triển của xã hội đòi hỏi hệ thống chăm sóc sức khỏe vận hành hiệu quả hơn, giảm chi phí và thời mà vẫn mang lại chất lượng cao. Thứ ba, sự bùng nổ dân số đã và đang diễn ra mạnh mẽ dẫn đến những thách thức như dịch bệnh bùng nổ, đòi hỏi sự kết hợp dữ liệu từ khắp nơi trên thế giới để cùng nhau giải quyết vấn đề mang tính toàn cầu trong đó có phân tích dữ liệu y khoa.

Trong những năm gần đây cùng với sự đột phá mạnh mẽ của việc ứng dụng công nghệ thông tin, công nghệ số và đặc biệt là trí tuệ nhân tạo vào phân tích dữ liệu lớn trong khoa học sức khỏe là hết sức cần thiết trong thời đại ngày nay. Phân tích ECG bằng trí tuệ nhân tạo là một nhánh nhỏ trong lĩnh vực phân tích dữ liệu y khoa, tập trung vào việc tăng độ chính xác của kết quả chẩn đoán và giảm tải công việc cho bác việc nhằm nâng cao hiệu suất khám và chữa bệnh.

Từ những vấn đề đặt ra như trên, cho thấy việc ứng dụng trí tuệ nhân tạo trong phân tích dữ liệu điện tâm đồ là hết sức cần thiết trong chẩn đoán các bệnh về tim nói riêng và xử lý dữ liệu y khoa nói chung.

II. MỤC TIÊU ĐỀ TÀI

Mục tiêu chính của đề tài là xây dựng và so sánh các mô hình máy học khác nhau để phân loại rối loạn nhịp tim.

Các mục tiêu cụ thể:

- Xây dựng các mô hình máy học phân loại rối loạn nhịp tim dựa trên dữ liệu điện tâm đồ.
- So sánh các mô hình máy học khác nhau.
- Tìm hiểu cách hoạt động của các mô hình máy học khác nhau.
- Tìm hiểu cách làm việc với dữ liệu chuỗi thời gian.

III. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

1. Đối tượng nghiên cứu

Đối tượng nghiên cứu chính của đề tài là các mô hình máy học khác nhau trong việc hỗ trợ phân loại rối loạn nhịp tim, từ đó so sánh hiệu suất, độ chính xác và khả năng áp dụng vào thực tế của các mô hình.

2. Phạm vi nghiên cứu

Phạm vi nghiên cứu tập trong vào việc tìm hiểu các thuật toán ở mức độ ứng dụng vào xử lý dữ liệu điện tâm đồ để tìm ra những bất thường, trong phạm vi niên luận không tập trung vào việc phân tích sâu các thuật toán máy học.

IV. PHƯƠNG PHÁP NGHIÊN CỨU

1. Nghiên cứu tài liệu

Nghiên cứu tài liệu gồm:

- Nghiên cứu các mô hình máy học chủ yếu dựa vào quyển *Hands-On Machine Learning with Scikit-Learn & TensorFlow.*
- Nghiên cứu tài liệu thư viện hỗ trợ gồm SciPy, NeuroKit2, Scikit-learn, PyTorch, ONNX,…
- Tài liệu về điện tâm đồ, rối loạn nhịp tim,…
- Các tài liệu có liên quan khác.

2. Thực nghiệm

Nghiên cứu thực nghiệm gồm:

- Thử nghiệm với dữ liệu điện tâm đồ.
- Xây dựng các mô hình máy học khác nhau dựa trên các thư viện.
- So sánh kết quả các mô hình đạt được.

V. NỘI DUNG NGHIÊN CỨU

Nội dung nghiên cứu tập trung vào việc xây dựng các mô hình máy học bằng các thư viện nguồn mở, đồng thời xử lý tín hiệu điện tâm đồ để nâng cao hiệu suất đào tạo và dự đoán của các mô hình.

VI. BỐ CỤC QUYỂN NIÊN LUẬN

Bố cục quyển niên luận gồm 3 phần chính như sau:

1. Phần Giới thiệu

Bao gồm các nội dung:

- Đặt vấn đề
- Mục tiêu đề tài
- Đối tượng và phạm vi nghiên cứu
- Phương pháp nghiên cứu
- Nội dung nghiên cứu
- Bố cục quyển niên luận

2. Phần Nội dung

Bao gồm các nội dung:

- Chương 1: Mô tả bài toán
- Chương 2: Cơ sở lý thuyết
- Chương 3: Mô tả tập dữ liệu
- Chương 4: Kết quả thực hiện
- Chương 5: Đánh giá kiểm thử

3. Phần Kết luận

Bao gồm các nội dung:

- Kết luận
- Hướng phát triển

PHẦN 2. NỘI DUNG
CHƯƠNG 1. MÔ TẢ BÀI TOÁN

1. Rối loạn nhịp tim

Theo Viện tim mạch Việt Nam thì ở người trưởng thành khỏe mạnh, nhịp đập bình thường của tim sẽ nằm trong khoảng 60-100 nhịp/phút lúc nghỉ ngơi. Rối loạn nhịp tim là tình trạng tốc độ hay nhịp đập của tim bất thường. Điều đó nghĩa là nhịp tim không ổn định, tim bạn có thể đập quá nhanh, quá chậm hoặc có nhịp tim không đều [1]. Rối loạn nhịp tim xảy ra do nhiều nguyên nhân chẳng hạn như căng thẳng, làm việc quá sức; sử dụng các chất kích thích như rượu, bia, thuốc lá,… ngoài ra rối loạn nhịp tim cũng có thể liên quan đến các bệnh lý như viêm cơ tim, hở van tim, thiếu máu cơ tim,… Bên cạnh đó, việc tăng huyết áp kéo dài, cường giáp, các bệnh liên quan tới phổi hoặc tác dụng phụ của thuốc cũng có thể gây rối loạn nhịp tim. Một số triệu chứng phổ biến của rối loạn nhịp tim là hồi hợp, đánh trống ngực, chóng mặt, đau ngực, khó thở, mệt mỏi,… khi xuất hiện những triệu chứng kể trên nên sớm gặp bác sĩ để tránh các biến chứng không mong muốn.

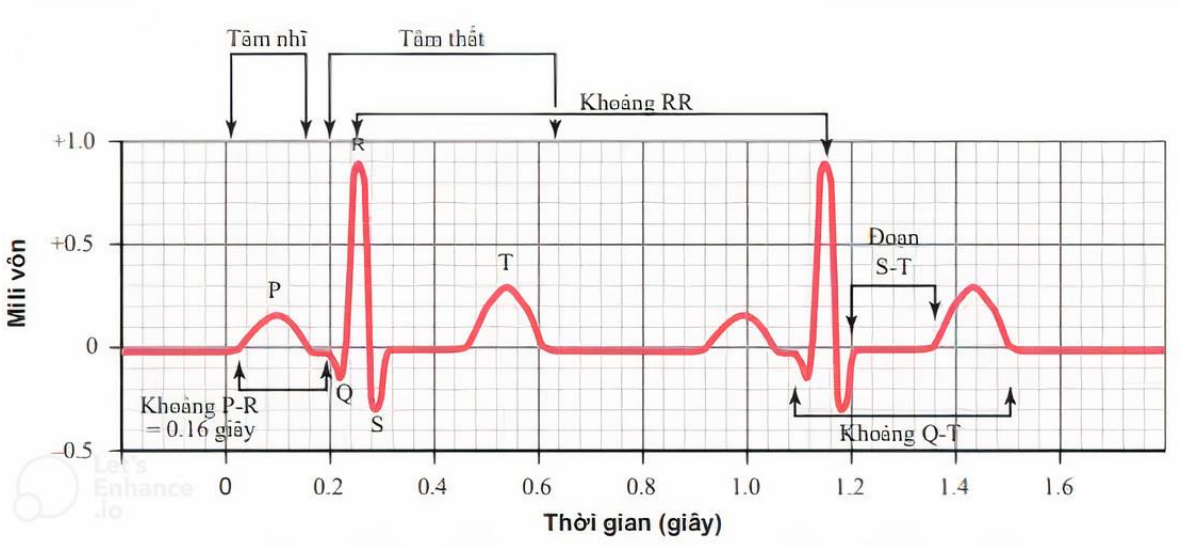
Một số loại rối loạn nhịp được sử dụng trong phạm vi niên luận gồm có:

- Rối loạn nhịp xoang (Sinus Rhythm) là tình trạng nhịp tim đập không đều, quá nhanh hoặc quá chậm.
- Rung nhĩ (Atrial Fibrillation) là tình trạng nhịp tim đập không đều, tín hiệu điện tim bị gián đoạn thường xảy ra ở người lớn tuổi, tình trạng này kéo dài có thể gia tăng nguy cơ đột quỵ.
- Block nhĩ thất (Atrioventricular Block) là bệnh lý tắc nghẽn đường dẫn truyền từ tâm nhĩ xuống tâm thất, được phát hiện bằng sự kéo dài khoảng PR trên điện tâm đồ. Nguyên nhân chính gây ra block nhĩ thất là do bệnh tim thiếu máu cục bộ hoặc xơ hóa vô căn và xơ cứng hệ thống dẫn truyền.
- Block nhánh trái/Block nhánh phải (Left Bundle Branch Block/Right Bundle Branch Block) là sự gián đoạn một phần hoặc toàn bộ dẫn truyền tín hiệu điện tim tại nhánh bên trái (hoặc bên phải) của hệ thống dẫn truyền điện tim sau khi đi ra từ bó His.
- Ngoại tâm thu nhĩ (Premature Atrial Contraction) là tình trạng nhịp nhanh trên thất hoặc nhịp nhanh kịch phát trên thất làm cho nhịp tim nhanh hoặc chậm hơn bình thường.
- Đoạn ST chênh lên/ Đoạn ST chênh xuống (ST Segment Depression/ ST Segment Elevation) là một dạng nhồi máu cơ tim với đặc trưng là đoạn ST chênh lên/chênh xuống trên điện tâm đồ.

2. Điện tâm đồ

Hoạt động co bóp của tim được điều khiển bởi hệ thống dẫn truyền gồm mạng lưới các nút, tế bào và tín hiệu điều khiển. Mỗi lần đập các tín hiệu điện được truyền qua tim khiến cho các bộ phận của tim giãn ra và co lại. Các tín hiệu điện này thường rất nhỏ (khoảng một phần nghìn volt) nhưng có thể được ghi lại bằng các điện cực đặt trên tay, chân và ngực bệnh nhân; sau đó được khuếch đại và ghi lại bằng điện tâm đồ.

Điện tâm đồ là đồ thị ghi những thay đổi của dòng điện trong tim theo thời gian [2]. Những dòng điện đó thường có hiệu điện thế rất nhỏ từ 1mV-3mV. Hình 1 minh họa điện tâm đồ của người bình thường. Khi khảo sát điện tâm đồ cần chú ý 9 đặc điểm: tần số và sự điều đặn, nhịp, sóng P, khoảng PR, phức bộ QRS, đoạn ST, sóng T, sóng U, khoảng QTc [3].



Hình 1: ECG nhịp tim bình thường

2.1. Sóng P – Nhĩ đồ

Hình 2: Sóng P

Xung động đi từ nút xoang (ở nhĩ phải) sẽ tỏa ra làm khử cực cơ nhĩ như các hình đợt sóng với hướng chung là từ trên xuống dưới và từ phải sang trái. Như vậy vectơ khử cực nhĩ sẽ có hướng từ trên xuống dưới và từ phải sang trái, làm với đường

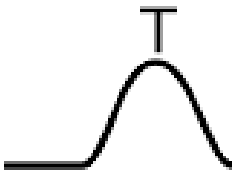
ngang một góc +49^{o} và còn gọi là trục điện nhĩ, tạo được một làn sóng dương thấp, nhỏ với thời gian khoảng 0,05 giây đến 0,1 giây gọi là sóng P. Do đó, trục điện nhĩ lại còn có tên là trục sóng P [3]. Nhĩ đồ là sự hoạt động của nhĩ chỉ thể hiện lên điện tim bằng một làn sóng đơn độc có hình dạng trơn láng như Hình 2: sóng P.

2.2. Thất đồ

Thất đồ được chia thành hai giai đoạn: giai đoạn khử cực gồm phức hợp QRS và được gọi là pha đầu, giai đoạn tiếp theo là tái cực bao gồm ST và T còn được gọi là pha cuối.

Hình 3: Phức hợp QRS

Sóng phức hợp QRS – khử cực xảy ra khi nhĩ đang khử cực rồi bắt đầu vào nút nhĩ – thất truyền qua thất và hai nhánh bó His xướng khử cực thất [4]. Khử cực thất gồm có ba làn sóng cao và nhọn là sóng Q, sóng R và sóng S biến thiên phức tạp nên được gọi là phức hợp QRS, trong đó sóng lớn nhất chính là sóng R minh họa như Hình 3.



Hình 4: Sóng T

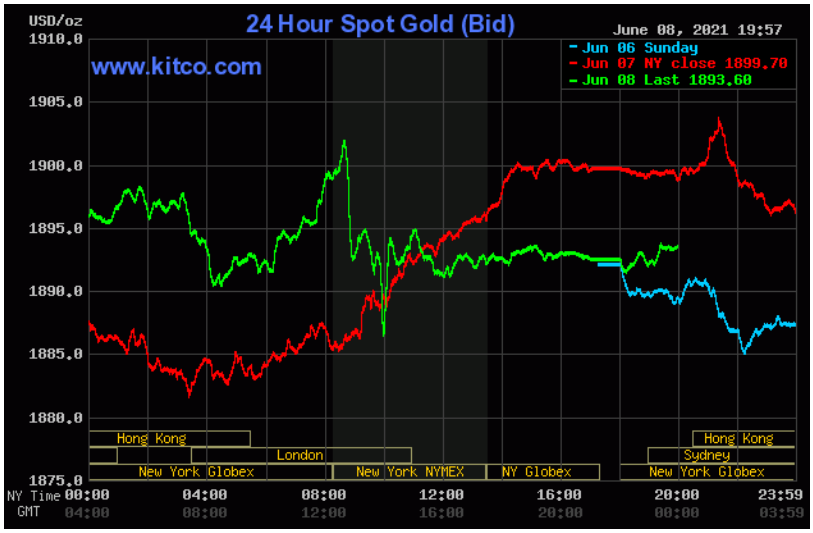
Sóng T – tái cực xảy ra trong thời kỳ tái cực chậm, thể hiện bằng một đoạn thẳng đồng điện gọi là đoạn ST trên điện tâm đồ, sau đó đến thời kỳ tái cực nhanh. Sóng T không đối xứng mà có sườn lên thoai thoải hơn và sườn xuống dốc đúng hơn. Thời gian của sóng T rất dài nên còn được gọi là sóng chậm minh họa như Hình 4.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

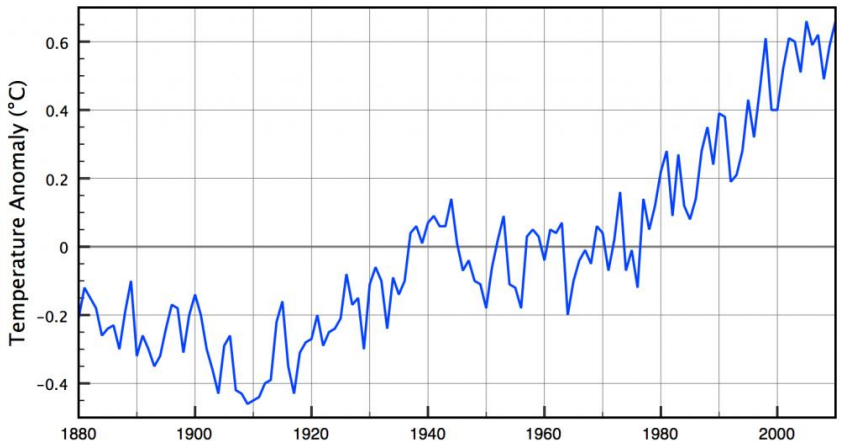
1. Dữ liệu chuỗi thời gian

Chuỗi thời gian là tập hợp các giá trị được đo đạc một cách tuần tự theo thời gian. Dữ liệu chuỗi thời gian được sử dụng trong nhiều lĩnh vực khác nhau như kinh tế, thời tiết, xã hội, y học,… Một số chuỗi thời gian thường thấy trong cuộc sống hàng ngày là chuỗi thời gian thể hiện lượng mưa theo từng tháng, giá chứng khoán hàng ngày, điện tâm đồ,… Hình 5 minh họa ví dụ về chuỗi thời gian giá vàng thế giới ngày 08 tháng 6 năm 2021, Hình 6 thể hiện ví dụ chuỗi thời gian nhiệt độ từ năm 1880 đến năm 2010.

Một số bài toán điển hình trong khai phá dữ liệu chuỗi thời gian bao gồm: Gom cụm (Clustering), Phân lớp (Classification), Phát hiện Motif (Motif detection), phát hiện chuỗi bất thường (Anomaly detection),…



Hình 5: Đường biểu diễn chuỗi thời gian giá vàng



Hình 6: Chuỗi thời gian nhiệt độ

2. Mô hình

Trong lĩnh vực trí tuệ nhân tạo nói chung và máy học nói riêng thì mục tiêu chính là xây dựng các chương trình được đào tạo để xử lý dữ liệu một cách tự động. Các mô hình này có thể được xây dựng bằng những thuật toán khác nhau hoặc kết hợp các thuật toán với nhau. Phần nội dung bên dưới trình bày các mô hình từ đơn giản đến phức tạp để xử lý dữ liệu điện tâm đồ tìm ra những bất thường.

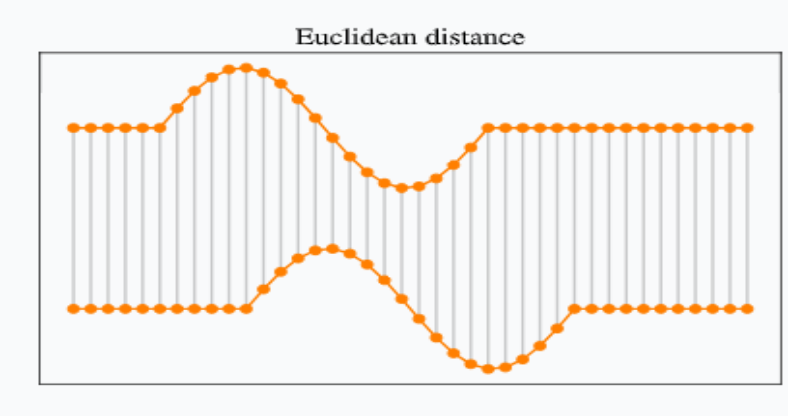
2.1. Baselines

Đường cơ sở gồm các mô hình máy học đơn giản dùng để trình bày ý tưởng giải quyết bài toán đặt ra, đồng thời so sánh với các mô hình máy học phức tạp hơn để có thể thấy được sự khác nhau và hiệu suất khi phân tích dữ liệu.

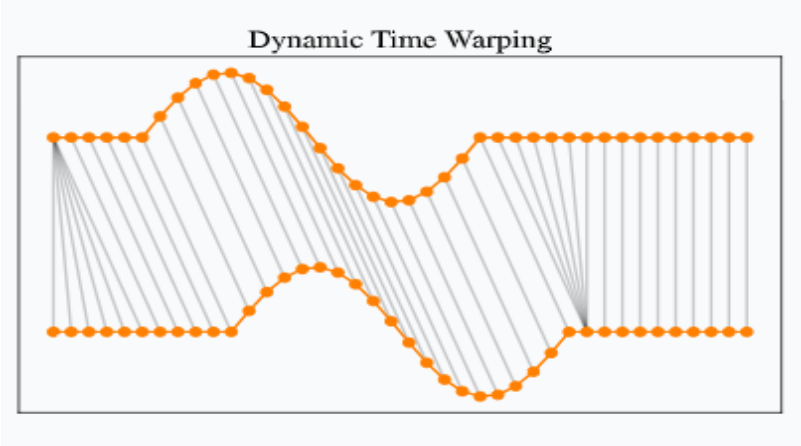
2.1.1. k-Nearest Neighbors with dynamic time warping

k-Nearest Neighbors (KNN) là một thuật toán nằm trong nhóm Supervised Learning, ý tưởng chính của thuật toán là đưa ra dự đoán dựa vào khoảng cách của nó và *k* láng giềng gần nó nhất. Khi đào tạo mô hình, thuật toán không làm gì cho đến khi có phần tử mới đến cần dự đoán thì mới tiến hành tính toán do đó còn được gọi xếp vào loại lazy learning. Một số khoảng cách được sử dụng trong KNN như khoảng cách Minkowski, khoảng cách Euclid, độ tương quan hoặc dynamic time warping.

Dynamic time warping (DTW) là một kỹ thuật nổi tiếng dùng để tìm sự liên kết tối ưu giữa hai chuỗi thời gian theo một số hạn chế nhất định [1]. Trong phân loại rối loạn nhịp tim bằng ECG vấn đề chính là phân giải quyết bài toán phân loại chuỗi thời gian (Time series classification) do đó việc áp dụng KNN và DTW là để so sánh mức độ giống nhau của hai chuỗi thời gian. Hình 7 minh họa sự khoảng cách Euclidean, Hình 8 minh họa DTW.



Hình 7: Khoảng cách Euclidean và sự biến dạng thời gian



Hình 8: Sự biến dạng thời gian

2.1.2. Logistic regression

Logistic Regression là thuật toán nằm trong nhóm Supervised Learning thường được dùng trong các bài toán phân loại nhị phân. Phương pháp Logistic Regression có thể được mở rộng để giải các toán phân loại nhiều lớp, phương pháp này thường được gọi là Softmax Regression hay Multinomial Logistic Regression.

Logistic Regression truyền thống:

Logistic Regression model estimated probability

$$\hat{p} = h_\theta(x) = \sigma(x^T\theta)$$

Logistic function (sigmoid function)

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Logistic Regression model prediction

$$y = \begin{cases} 0 \; if \; \hat{p} < 0.5 \\ 1 \; if \; \hat{p} \geq 0.5 \end{cases}$$

Cost function of a single training instance

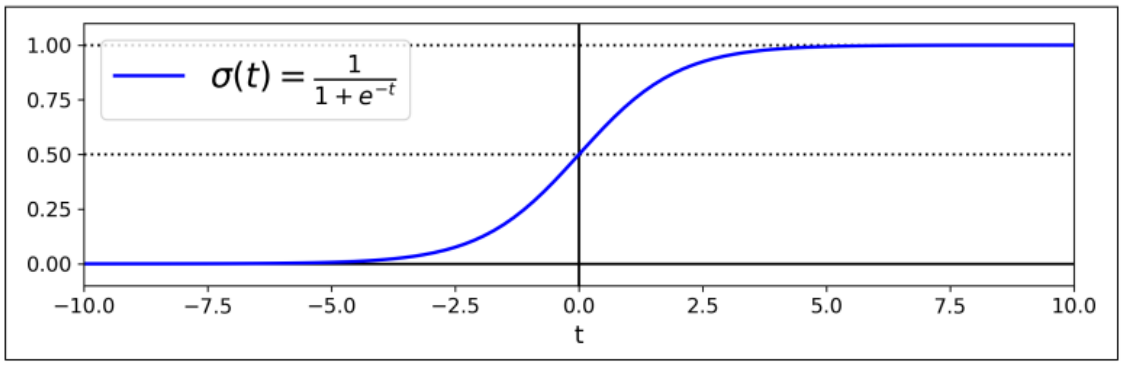
$$c(\theta) = \begin{cases} -\log(\hat{p}) \; nếu \; y = 1 \\ -\log(1 - \hat{p}) \; nếu \; y = 0 \end{cases}$$

Logistic Regression cost function (log loss)

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log(\hat{p}^{(i)}) + (1 - y^{(i)})\log(1 - \hat{p}^{(i)})]$$

Logistic cost function partial derivatives

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(\sigma(\theta^T x^{(i)}) - y^{(i)})x_j^{(i)}$$



Hình 9: Sigmoid function

Softmax Regression

Softmax score for class k

$$s_k(x) = x^T \theta^{(k)}$$

Softmax function

$$\hat{p}_k = \sigma(s(x))_k = \frac{e(s_k(x))}{\sum_{j=1}^{K} e(s_j(x))}$$

- *K is the number of classes*
- *s(x) is a vector containing the scores of each class for instance*
- $\sigma(s(x))_k$ is the estimated probability that the instance x below the scores of each class for that instance

Softmax Regression classifier prediction

$$\hat{y} = \underset{k}{argmax}\,\sigma(s(x))_k = \underset{k}{argmax}\,s_k(x) = \underset{k}{argmax}\,((\theta^{(k)})^T x)$$

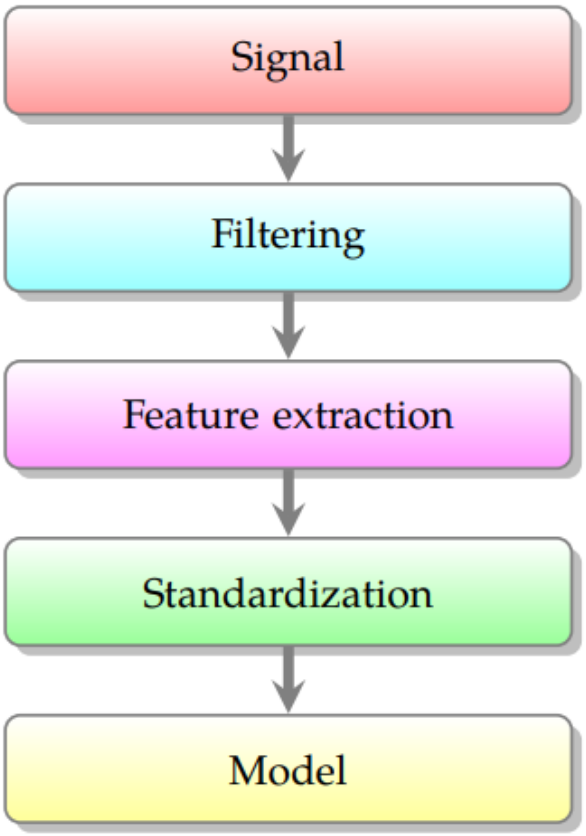
Cross entropy cost function

$$J = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log(\hat{p}_k^{(i)})$$

Công thức được trích từ [7, pp. 144-151]

2.2. Models with hand-engineered features

Features Engineered là quá trình chuyển đổi tập dữ liệu thô ban đầu thành tập các thuộc tính [8]. Phương pháp này sử dụng kiến thức miền (domain knowledge) để trích xuất các features đặc trưng của tín hiệu điện tim được tóm tắt như Hình 10.



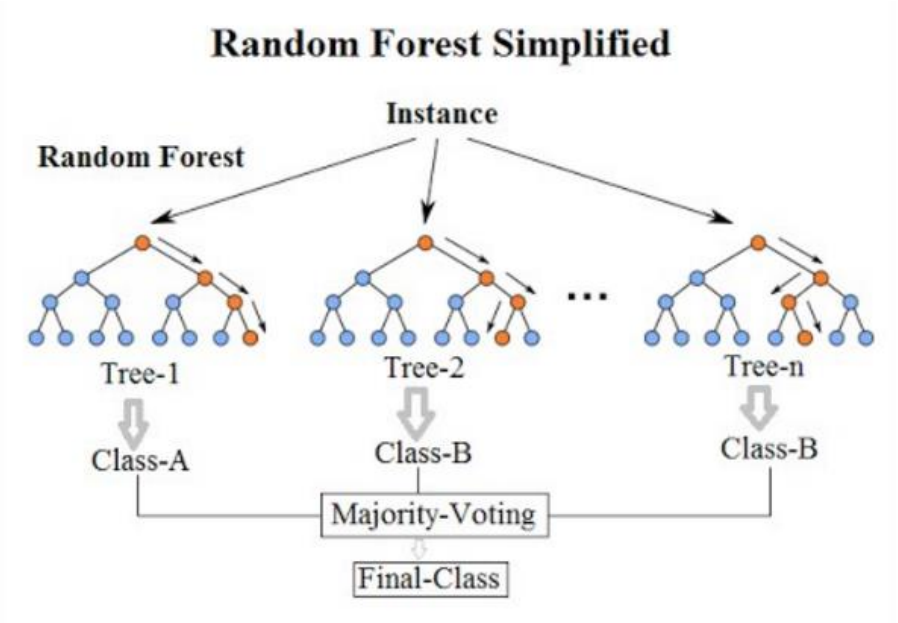
Hình 10: Models with hand-engineered features

2.2.1. Random forest

Bootstrap Aggregating (hay còn gọi là Bagging) là phương pháp xây dựng mô hình bằng cách kết hợp các mô hình cơ sở độc lập với nhau nhằm giảm lỗi xảy ra do tính biến thiên của mô hình so với tính ngẫu nhiên các mẫu dữ liệu của tập dữ liệu (lỗi variance). Các mô hình cơ sở này thường cùng loại với nhau nhưng được xây dựng trên những mẫu khác nhau của tập dữ liệu. Từ bộ dữ liệu ban đầu tiến hành lấy mẫu có hoàn lại để sinh ra k bộ dữ liệu mới.

Decision Tree là một thuật toán thuộc nhóm Supervised Learning có thể giải quyết cả bài toán hồi quy và phân lớp. Ý tưởng chính của thuật toán là xây dựng cây phân cấp dựa trên các luật để đưa ra các dự đoán. Trong thuật toán Decision Tree có thể gặp phải vấn đề là tỷ lệ phân loại đúng trên tập dữ liệu đào tạo cao nhưng tỷ lệ dự đoán trên tập kiểm tra lại thấp do độ sâu tùy ý, dẫn tới lỗi variance cao.

Random forest là phương pháp thuộc nhóm Ensemble model sử dụng mô hình cơ sở là Decision Tree ra đời nhằm giải quyết vấn đề lỗi variance cao. Các mô hình cơ sở này được xây dựng một cách độc lập trên tập mẫu Bootstrap, nằm trong nhóm Bagging. Thuật toán Random Forest gồm nhiều cây quyết định do đó kết quả dự đoán dựa trên luật bình chọn số đông đối với bài toán phân loại và dựa trên giá trị trung bình của các mô hình cơ sở đối với bài toán hồi quy, Hình 11 minh họa kết quả dự đoán của bài toán phân loại.



Hình 11: Minh họa thuật toán Random Forest Simplified

2.2.2. XGBoost

Gradient Boosting là một thuật toán xây dựng tập hợp các mô hình một cách tuần tự, mô hình sau sẽ học học cách để sửa lỗi của mô hình trước [7, p. 205]. Mục tiêu chính của boosting là giảm lỗi liên quan đến mô hình (bias), vấn đề đặt ra là xây dựng thuật toán nhằm giải quyết bài toán tối ưu.

Bài toán đặt ra

$$min_{c_n=1:N, w_n=1:N} L\left(y, \sum_{n=1}^{N} c_n w_n\right)$$

Trong đó:

L: giá trị hàm mất mác

y: nhãn

C_n: *trọng số*

w_n: *mô hình học yếu thứ n*

Một số đặc điểm của Boosting:

- Thời gian đào tạo mô hình tương đối lâu do các mô hình được xây dựng một cách tuần tự.
- Sau mỗi vòng lặp, Boosting có thể giảm lỗi theo cấp số nhân.
- Boosting sẽ hoạt động tốt nếu mô hình cơ sở không quá phức tạp.
- Boosting có thể giảm lỗi bias cho các mô hình cơ sở.

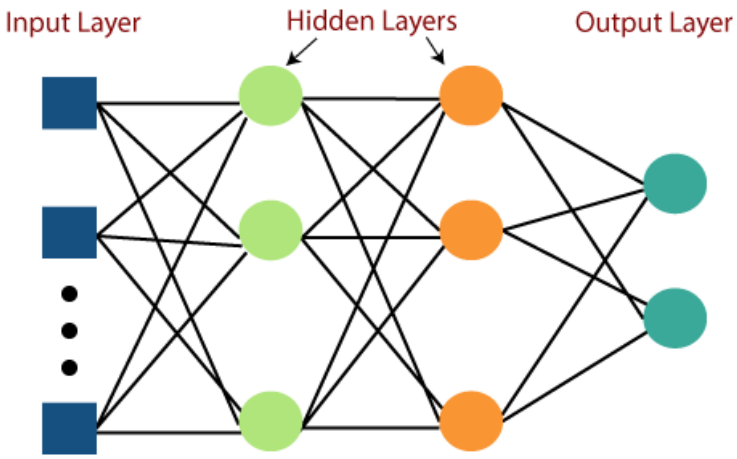
Extreme Gradient Boosting (XGBoost) là một giải thuật học dựa trên Gradient Boosting, tuy nhiên kèm theo đó là những cải tiến to lớn về mặt tối ưu thuật toán, về sự kết hợp hoàn hảo giữa sức mạnh phần mềm và phần cứng, giúp đạt được những kết quả vượt trội cả về thời gian đào tạo mô hình cũng như bộ nhớ sử dụng [3]. XGBoost hỗ trợ nhiều nền tảng, nhiều hệ sinh thái khác nhau và có khả năng giải quyết bài toán hồi quy lẫn phân lớp. Một số điểm mạnh của XGBoost so với Gradient Boosting:

- Có khả năng tránh overfitting.
- Tận dụng được tài nguyên hệ thống để tiến hành tính toán song song.
- Khả năng missing data value để giảm thời gian đào tạo mô hình.

Trong phần niên luận có sử dụng XGBoostClassifier từ thư viện Scikit-learn, mô hình cơ sở được sử dụng là Decision Tree.

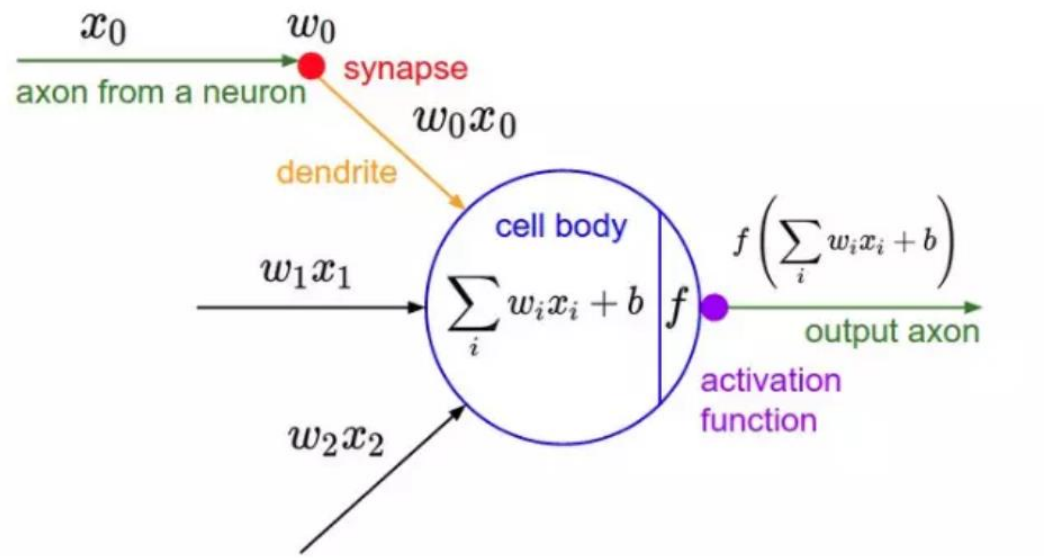
2.2.3. Multilayer perceptron

Multilayer perceptron (MLP) là một lĩnh vực nghiên cứu về mạng thần kinh nhân tạo. Hình 12 minh họa một mạng nơ-ron nhân tạo gồm 3 tầng: tầng vào (input layer), tầng ẩn (hidden layer) và tầng ra (output layer); mỗi tầng gồm một hay nhiều neural liên kết với các neural của tầng liền kề với nó.



Hình 12: Mạng neural đa tầng

Hình 12 minh họa một neural nhân tạo (hay còn được gọi là perceptron) được đề xuất bởi McCulloch và Pitts.



Hình 13: Tương quan giữa neural sinh học và neural nhân tạo

Hàm mạng tuyến tính:

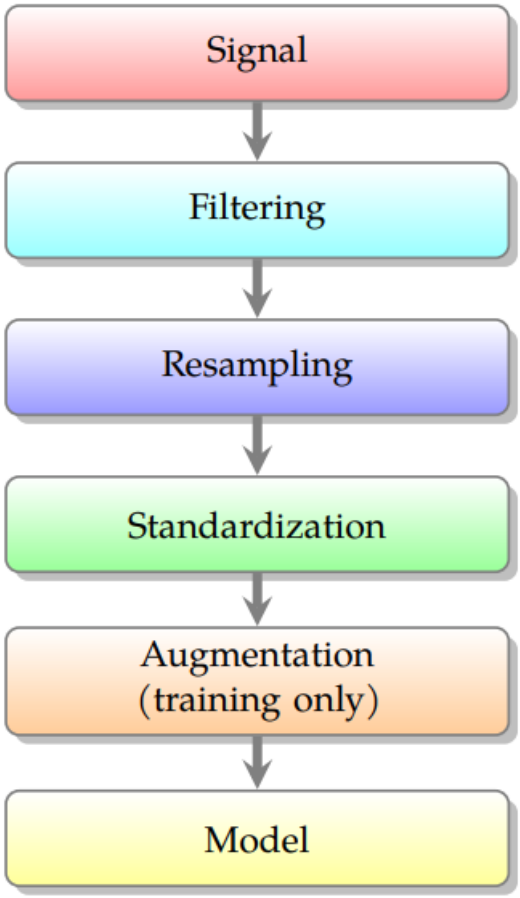
$$u = g(x) = \sum_{i=0}^{n} w_i x_i$$

Hàm kích hoạt:

$$o = f(u) = f(g(x))$$

2.3. Models with direct signal input

Khác với phương pháp Feature Engineered đã trình bày bên trên, phương pháp direct signal input không cần nhiều domain knowledge. Các mô hình có nhiệm vụ chính là trích xuất các tính năng, ánh xạ giữa các mẫu và nhãn. Hình 14 trình bày khái quát mô hình direct signal input.



Hình 14: Models with direct signal input

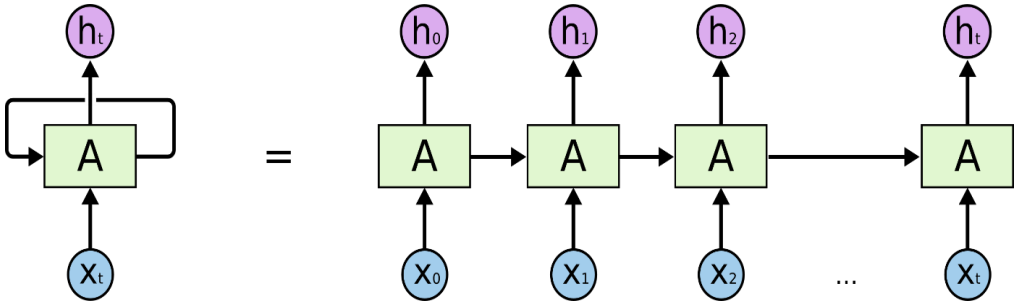
2.3.1. ResNet

Convolutional networks là một loại mạng nơ-ron nhân tạo bao gồm các lớp phức tạp. Các perceptron không được kết nối dày đặc như MLP, thay vào đó đầu vào của neural chỉ từ một vùng nhỏ trước đó [10]

ResNet (Residual network) là một dạng của mạng neural tích chập ra đời năm 2015 trong lĩnh vực thị giác máy tính. ResNet đào tạo mô hình với độ sâu lớn mà không làm giảm hiệu suất của mô hình học sâu. Trong phạm vi niên luận, kiến trúc ResNet được sử dụng với 18 lớp tích chập để xây dựng mô hình.

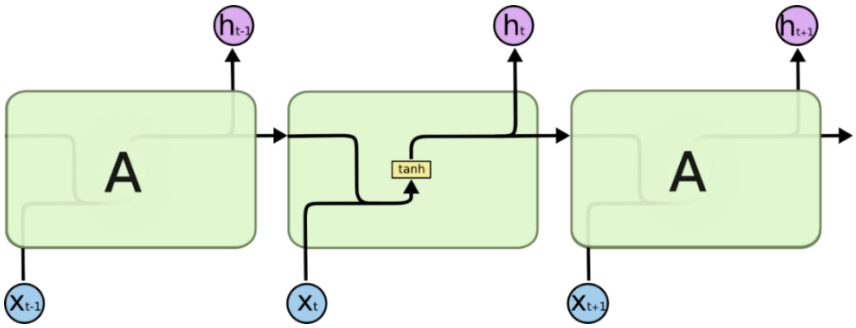
2.3.2. Recurrent networks

Recurrent neural networks (RNN) minh họa như Hình 15 gồm vòng lặp giúp lưu lại các thông tin trước đó. HÌNH mô tả diễn giải kiến trúc của mạng nơ-ron hồi quy A với đầu vào x_t và đầu ra h_t, về cơ bản thì kiến trúc RNN cũng tương tự với các mạng nơ ron truyền thống. RNN gồm một vòng lặp cho phép các thông tin có thể được truyền từ bước này qua bước khác của mạng nơ-ron, thông tin vừa là đầu ra của mạng này đồng thời là đầu vào của mạng khác.



Hình 15: Recurrent networks

RNN đều có dạng một chuỗi các mô-đun lặp đi lặp lại của mạng nơ-ron. Với mạng RNN chuẩn, các mô-đun này có cấu trúc rất đơn giản, thường là một hàm tanh. Hình 16 minh họa cấu trúc một RNN chuẩn. Ý tưởng chính của RNN chính là sử dụng lại những thông tin đã có trước đó để dự đoán hiện hiện tại, tương tự việc con người dùng những gì học được trong quá khứ nhằm quyết định tương lai.



Hình 16: Module của một RNN

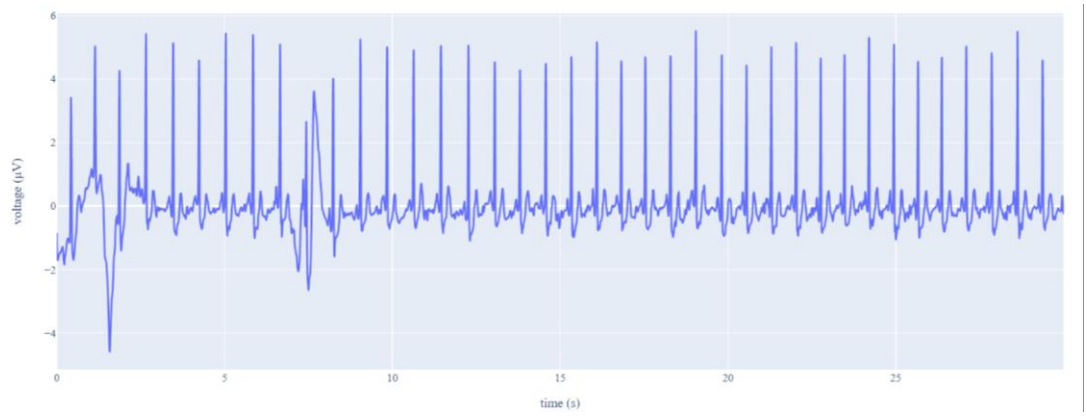
CHƯƠNG 3. MÔ TẢ TẬP DỮ LIỆU

1. The Physionet computing in Cardiology challenge 2017

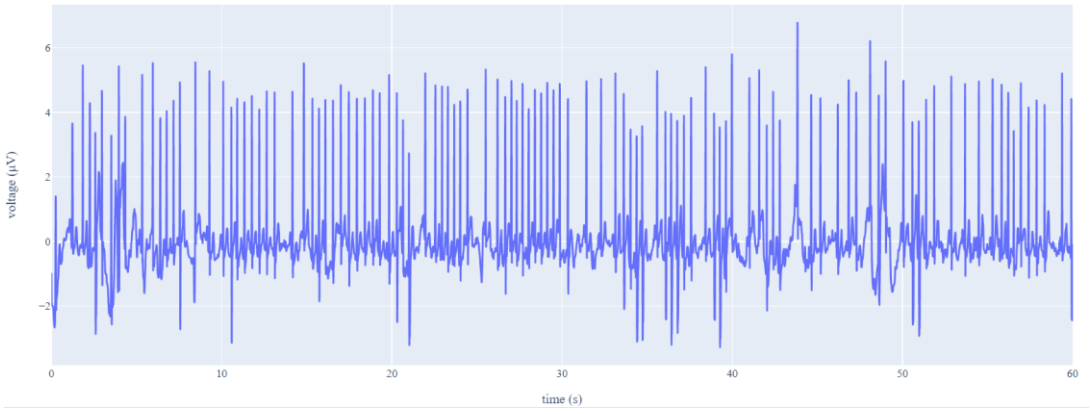
Bộ dữ liệu Cinc2017Dataset được sử dụng trong cuộc thi "AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017" [11]. Tổng cộng 12186 bản ghi điện tâm đồ được tài trợ bởi AliveCor được ghi lại bởi các thiết bị của AliveCor. Bộ dữ liệu đào tạo (training set) gồm 8528 bản ghi có độ dài từ 9 – 61 giây được lấy mẫu ở 300Hz, bộ dữ liệu kiểm tra (test set) gồm 3658 mẫu có độ dài tương tự. Bảng 1 mô tả nhãn và số lượng nhãn của bộ dữ liệu Cinc2017Dataset; Hình 17, Hình 18, Hình 19 và Hình 20 minh họa một số mẫu của bộ dữ liệu Cinc2017Dataset.

Bảng 1: Bộ dữ liệu Cinc2017Dataset

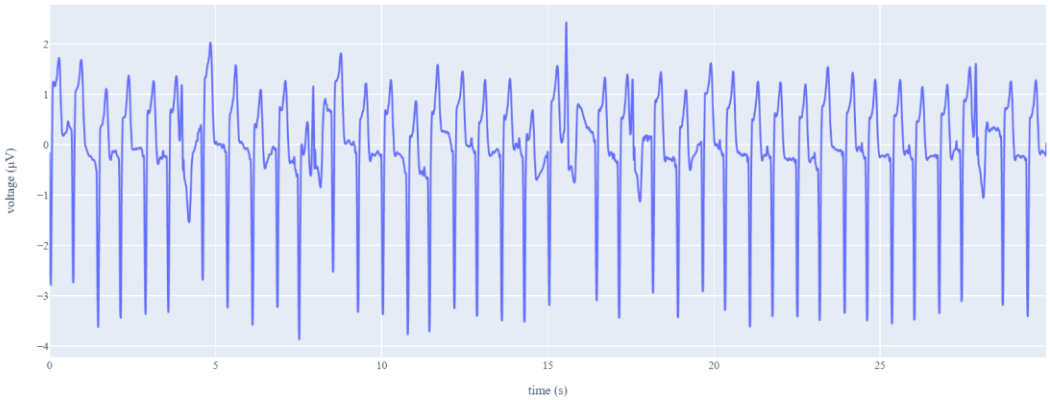
Label	Condition	Count
N	Sinus Rhythm	5154
A	Atrial Fibrillation	771
O	Other Rhythm	2557
X	Noisy	46
Total		8528



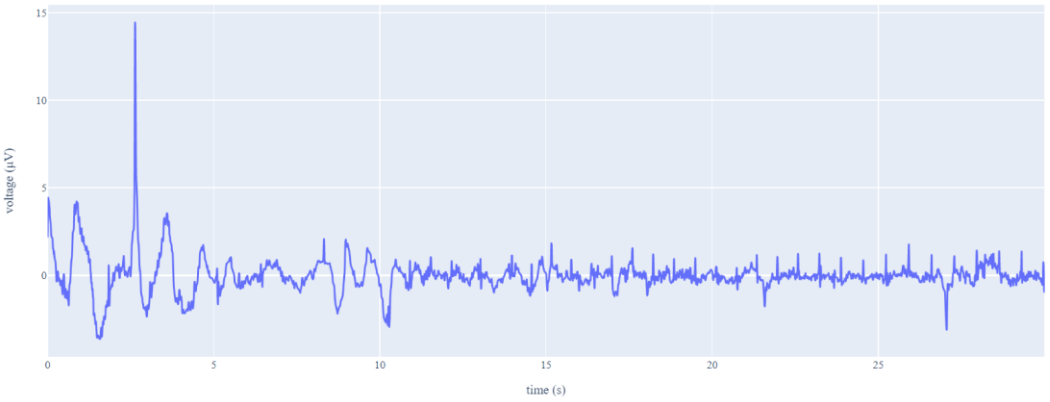
Hình 17: Rối loạn nhịp xoang – A00001



Hình 18: Rung nhĩ – A00005



Hình 19: Rối loạn nhịp tim khác – A08525



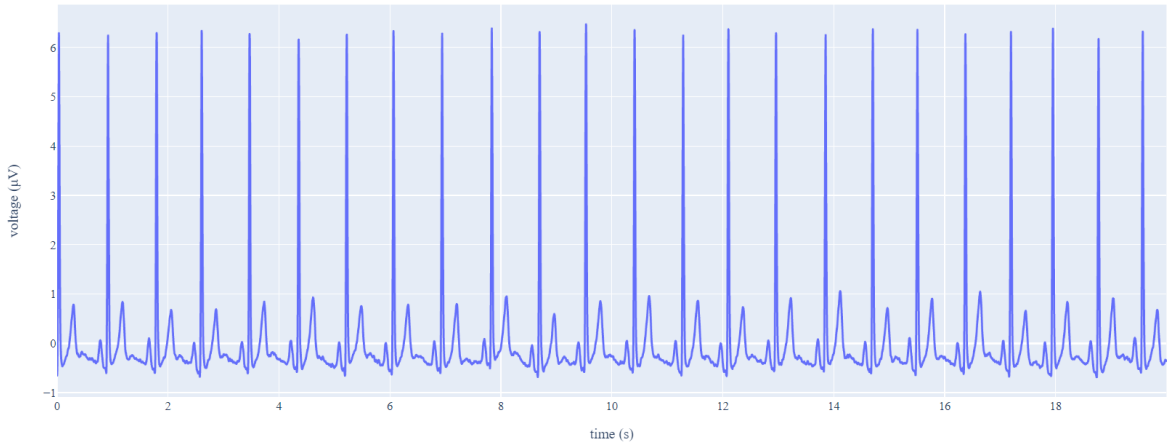
Hình 20: Tín hiệu bị nhiễu – A00022

2. The China Physiological Signal Challenge 2018

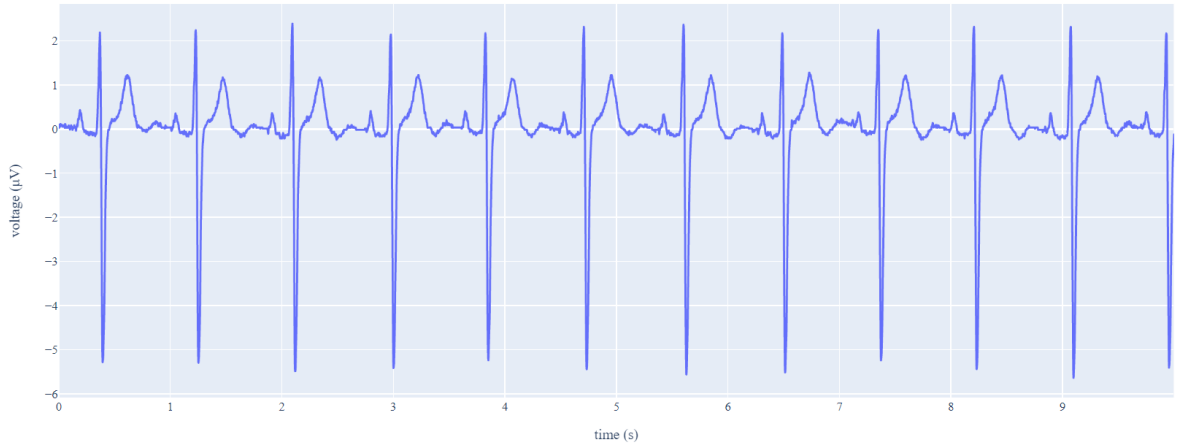
Bộ dữ liệu Cpsc2018Dataset được sử dụng trong cuộc thi "The China Physiological Signal Challenge 2018" [12]. Bộ dữ liệu có tổng cộng 9831 bản ghi lấy từ 9458 bệnh nhân khác nhau. Mỗi mẫu có độ dài từ 6-60 giây được lấy mẫu ở 500Hz. Điện tâm đồ 12 đạo trình được sử dụng trong bộ dữ liệu Cpsc2018Dataset gồm 1 nhãn bình thường và 8 nhãn bất thường được mô tả chi tiết như Bảng 2.

Bảng 2: Bộ dữ liệu Cpsc2018Dataset

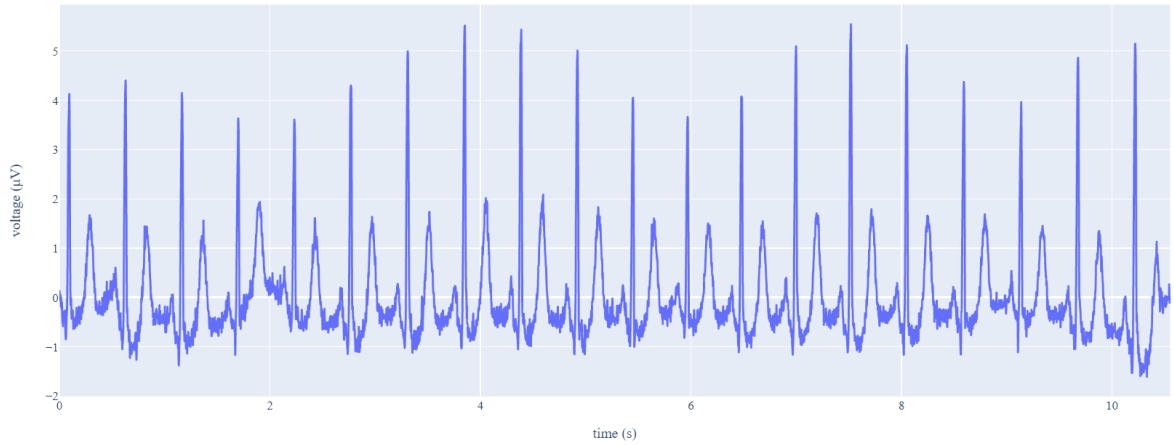
Label	Label code	Condition	Count
Normal	0	Sinus Rhythm	918
AF	1	Atrial fibrillation	876
I-AVB	2	First-degree Atrioventricular Block	686
LBBB	3	Left Bundle Brunch Block	179
RBBB	4	Right Bundle Brunch Block	1533
PAC	5	Premature Atrial Contraction Premature	532
PVC	6	Ventricular Contraction	607
STD	7	ST-Segment Depression	784
STE	8	ST-Segment Elavation	185
Total			6400



Hình 21: Nhịp tim bình thường – A00016



Hình 22: Block nhĩ thất – A00039



Hình 23: Đoạn ST chênh lên – A00033

CHƯƠNG 4. KẾT QUẢ THỰC HIỆN

I. PHÂN CHIA TẬP DỮ LIỆU

Tập dữ liệu được chia thành hai phần dùng để đào tạo (datatrain) và kiểm tra (datatest) theo tỷ lệ lần lượt là 90 – 10%, giống nhau cho tất cả mô hình. Trong quá trình đào tạo mô hình, training set được chia thành ba tập con là train, validation và test theo tỷ lệ lần lượt là 80-10-10%. Bảng 3 minh họa số lượng cụ thể phần tử của của tập dữ liệu đào tạo và kiểm tra.

Bảng 3: Train-test split

		Cinc2017Dataset	Cpsc2018Dataset
Test set		853	640
Training set	Train	6139	4608
	Validation	768	576
	Test	768	576

Cross-validation là một phương pháp dùng để đánh giá hiệu quả của các mô hình máy học, thường được sử dụng để so sánh và chọn lựa mô hình tốt nhất để giải quyết bài toán. Kỹ thuật này thường được sử dụng khi có ít dữ liệu; tham số k rất quan trọng trong kỹ thuật này do đó còn được gọi là k-fold cross validation. K-fold cross validation gồm các bước như sau:

1. Xáo trộn tập dữ liệu một cách ngẫu nhiên.
2. Chia tập dữ liệu thành k phần bằng nhau.
3. Với mỗi nhóm ta thực hiện các bước:
   - Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình.
   - Các nhóm còn lại dùng huấn luyện mô hình.
   - Huấn luyện.
   - Đánh giá và hủy mô hình.
4. Đánh giá hiệu suất mô hình.

Hold-out là một phương pháp dùng để đánh giá hiệu quả của các mô hình máy học, tập dữ liệu được chia thành hai phần riêng biệt không giao nhau là datatrain dùng để đào tạo mô hình và dataset dùng để kiểm thử mô hình.

Các mô hình trong phạm vi niên luận phân chia tập dữ liệu theo nghi thức k-fold cross-validation với k = 5 ngoại trừ CNN và RNN. Đối với CNN và RNN sử dụng nghi thức hold-out để đánh giá mô hình.

II. MODELS WITH HAND – ENGINEERED FEATURES

Tất cả mô hình thuộc nhóm models with hand-engineered features được đào tạo bằng cách sử dụng 5-fold cross-validation trong số 90% dữ liệu, các tham số được sử dụng để đào mô hình được trình bày như Bảng 4.

Bảng 4: Tham số các mô hình engineered features

Name	Parameter
Logistic Regression	The regularization parameter C = 100.0
Random forest	Max_depth = 20 Min_samples_leaf = 5 Number of estimators = 1000
XGBoost	Learning rate eta = 0.1 Gamma = 0.1 Max_depth of tree = 7 Min_child_weigth = 4 Number of estimators = 1000 Subsample parameter = 0.8
MLP	Learning rate = 0.0001 Batch size = 128

III. MODELS WITH DIRECT SIGNAL INPUT

ResNet được sử dụng từ thư viện Torchvision cho các xử lý liên quan đến thị giác máy tính. Cấu trúc mạng gồm 18 tầng (layer) và 50 lớp tích chập (convolutional layer) , gồm 20 epochs sau khi kết thúc mỗi epoch sẽ tiến hành đo xác thực. Tỷ lệ học được sử dụng là 0.0003 và trọng số phân rã là 0.0001.

CnnGru sử dụng tốc độ học là 0.0003 và trọng số phân rã là 0.0001, cross-entropy loss; gồm 10 epochs và lưu mô hình với điểm xác thực cao nhất.

CHƯƠNG 5. ĐÁNH GIÁ KIỂM THỬ

I. ĐÁNH GIÁ MÔ HÌNH

1. Confusion Matrix

Một mô hình máy học sau khi xây dựng cần một phép đo để đánh giá tính khả thi cũng như so sánh với các mô hình máy khác. Cách thường được sử dụng nhất là tính accuracy dựa trên tổng số lượng điểm dự đoán đúng trên tổng số lượng điểm có trong tập kiểm tra. Tuy nhiên, vấn đề đặt ra của bài toán phân lớp hay cụ thể là bài toán phát hiện rối loạn nhịp tim là phát hiện ra những bất thường trong tập dữ liệu. Do đó nếu sử dụng độ chính xác để đánh giá mô hình có thể dẫn đến việc độ chính xác rất cao nhưng không phát hiện được bất thường dẫn đến kết quả chẩn đoán có thể sai lệch nghiêm trọng. Confusion được sử dụng để đánh giá các mô hình trong phạm vi niên luận nhằm giải quyết vấn đề vừa nêu ra.

Confusion matrix đối với bài toán phân loại nhị phân gồm hai lớp là Negatives và Positives gồm hai hàng và hai cột trình bày như Bảng 5.

Bảng 5: Ma trận phân loại nhị phân

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (TN)

Ma trận gồm một số thuật ngữ:

- True Positives (TP): dương tính thật.
- False Positives (FP): dương tính giả.
- True Negatives (TP): âm tính thật.
- False Negatives (FN): âm tính giả.
- Accuracy: độ chính xác được tính theo công thức:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

- True Positive Rate - Recall (TPR): tỷ lệ dương tính thực (độ nhạy) được tính theo công thức:

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

- Positive Predictive Value – Precision (PPV): tỷ lệ dương tính đoán đúng được tính theo công thức:

$$PPV = \frac{TP}{TP + FP}$$

2. F1 score

F1 score là trung bình điều hòa (harmonic mean) của tỷ lệ dương tính thực và tỷ lệ dương tính đoán đúng, được tính theo công thức:

$$F1 = 2 * \frac{PPV * TPR}{PPV + TPR}$$

F1 score có giá trị nằm trong nửa khoảng (0; 1], đây là cách đơn giản để so sánh các bộ phận phân loại. F1 càng cao thì bộ phân lớp càng tốt và ngược lại.

II. KẾT QUẢ

Bảng 6 trình bày kết quả so sánh các mô hình sử dụng F1 score

Bảng 6: Kết quả mô hình sử dụng F1 score

ID	Model	Cinc2017Dataset	Cpsc2018Dataset
1	LogisticRegression	0.64353953	0.452028719
2	RandomForestClassifier	0.651264298	0.427094167
3	XGBClassifier	0.651414266	0.488561762
4	MLPClassifier	0.686923369	0.49091005
5	ResNet18Classifier	0.679418092	0.62858324
6	CnnGruClassifier	0.640398572	0.545206662

PHẦN 3. KẾT LUẬN

I. KẾT QUẢ ĐẠT ĐƯỢC

Một số kết quả đạt được:

- Xây dựng được các mô hình:
- Phát hiện được các bất thường trên dữ liệu điện tâm đồ.
- Nghiên cứu lý thuyết chuỗi thời gian.

II. HẠN CHẾ

Đề tài chỉ làm việc với từng bộ dữ liệu cụ thể, chưa thể làm việc với tất cả dữ liệu điện tâm đồ một cách độc lập.

Dữ liệu nghiên cứu tương đối hạn chế do đó chưa thể đánh giá được hết tính khả thi của mô hình khi áp dụng vào thực tế.

Kiến thức chuyên môn về lĩnh vực y khoa là một khó khăn lớn vì có những kiến thức chuyên sâu khó tiếp cận.

III. HƯỚNG PHÁT TRIỂN

Tiếp tục cải thiện chất lượng dự đoán của mô hình, tăng độ chính xác và giảm thời gian huấn luyện.

Bổ sung thêm một số kiến thức chuyên môn về lĩnh vực điện tim để nâng cao khả năng thực tiễn của mô hình.

TÀI LIỆU THAM KHẢO

[1] N. Anh, "hellobacsi," [Online]. Available: https://hellobacsi.com/benh-tim-mach/roi-loan-nhip-tim/roi-loan-nhip-tim/. [Accessed 17 October 2022].

[2] Đặng Xuân Thắng, Phạm Đức Hùng, "Công nghê Công nghệ cao," [Online]. Available: https://congnghiepcongnghecao.com.vn/tin-tuc/t24416/ung-dung-tri-tue-nhan-tao-trong-tim-mach-hoc.html. [Accessed 17 October 2022].

[3] P. N. Vinh, Sổ tay điện tâm đồ, Nhà xuất bản y học, 2018.

[4] T. Đ. Trinh, Hướng dẫn đọc điện tâm đồ, Huế: Đại học Y Được Huế, 2008.

[5] N. V. H. Anh, Tách phức hợp QRS bằng phương pháp Wavelet, Cần Thơ: Trường Đại học Cần Thơ, 2013.

[6] M. Muller, Information Retrieval for Music and Motion, Springer, January 2007.

[7] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd ed., O'Reilly Media, 2019.

[8] Ô. X. Hồng, "Ông Xuân Hồng - Chia sẻ kiến thức và thông tin về Machine Learning," 29 October 2015. [Online]. Available: https://ongxuanhong.wordpress.com/2015/10/29/feature-engineering-la-gi/. [Accessed 10 November 2022].

[9] B. T. Tùng, "https://viblo.asia," 28 May 2021. [Online]. Available: https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0.

[10] A. Ivora, *ECG Arrhythmia Detection and Classification,* Brno, 2020, p. 24.

[11] "PhysioNet," MIT Laboratory for Computational Physiology, 1 February 2017. [Online]. Available: https://physionet.org/content/challenge-2017/1.0.0/#files-panel. [Accessed 27 December 2022].

[12] [Online]. Available: http://2018.icbeb.org/Challenge.html.

[13] 8 September 2022. [Online]. Available: https://en.wikipedia.org/wiki/Multilayer_perceptron.