國立臺灣大學工學院應用力學研究所

碩士論文

Institute of Applied Mechanics

College of Engineering

National Taiwan University

Master's thesis

# 藉由機器學習判別伴隨心房顫動之中風病患的嚴重程度

# Outcome Prediction in Stroke Patients with Atrial Fibrillation using Machine Learning Techniques

郭昱德

Yu-te Kuo

指導教授: 潘斯文 博士

Advisor: Stephen Payne,DPhil.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202403578

# Acknowledgements

I am deeply honored to be involved in a clinical-oriented study aimed at establishing an AI model for diagnosis. My sincere gratitude goes to the doctors and staff at the National Taiwan University Hospital (NTUH), as well as to my advisor who has been instrumental in bridging the gap between physicians and myself, given our different backgrounds and knowledge bases. He played a critical role to connect the engineering and medical professionals since biomedical engineering is a novel and multi-discipline field to explore. Additionally, I would like to thank them for their great aid in providing data for scholarly research.

Although our approach to medical study may differ from the traditional methods employed by clinical doctors, we hold the utmost respect for the diagnostic expertise of medical professionals. Their specialised training, background knowledge, and unwavering commitment to pushing the boundaries of medical research contribute significantly to the advancement of clinical knowledge. Furthermore, their dedication extends to the dissemination of health education to the global community, enhancing public awareness and understanding of medical matters.



doi:10.6342/NTU202403578

# 中文摘要

心律不整是一個當今老化的社會中盛行的心血管疾病其中一個嚴峻的課題，嚴重的情況可能導致死亡。其中最常見型態的心律不整為心房顫動。

房顫是由於心臟組織失調的電位活動，導致心臟異常的收縮甚至產生顫動的情形。然而，有關於中風相關的疾病背後的病理學牽涉複雜，因此我們排除其他非缺血性中風的其他病理學所導致的中風。

為了描述這個問題，我們採用由臨床內科醫師確診為心房顫動暨中風症狀的病患的來自加護病房生理訊號，探索使用機器學習的方法來判別病患的中風嚴重程度。

我們的目標是試圖探索那些特徵是有效足以讓我們能夠有效鑑別不同中風嚴重程度的病患。

我們研究的發現即時血壓分析之於彌補心率變異性分析（HRV）的重要性，特別是心電圖(ECG)的量測。如果沒有涉及到與收縮壓在時間和資訊熵域條件特徵，我們很難泛化各種背後複雜的病理學。此外，光體積描記法（PPG）對於研究血流速率和心血管健康也是至關重要。

**關鍵字:缺血型中風;房顫;心律變異性分析;資訊熵;光體積變化描繪法;監督式學習**



doi:10.6342/NTU202403578

# Abstract

Arrhythmia is emerging as a significant cardiovascular disease in our modern aging society and can lead to fatal outcomes; One of the most prevalent types of arrhythmia is Atrial Fibrillation (AF). AF originates from abnormal discharges in cardiac tissue, resulting in incomplete atrial contraction and atrial fibrillation. To address this issue, we analysed the recordings of acute ischaemic stroke patients associated with AF symptoms diagnosed by clinicians for further analysis and explored the methodology to classify them into different severity of stroke by Machine Learning Technique. Our goal is to explore what features are significant for distinguishing the different outcomes of this patient group.

Our study findings underscore the critical role of real-time blood pressure analysis in complementing Heart Rate Variability (HRV), especially in ECG measurements. Without incorporating haemodynamic condition features related to Systolic blood pressure in both Time and Entropy domains, it's difficult to generalise the complex metrics underlying pathology. Additionally, Photoplethysmography (PPG) is indispensable for investigating flow rate and cardiovascular health.



doi:10.6342/NTU202403578

**Keywords: Ischaemic stroke; Atrial Fibrillation; Information Entropy;**

**Plethysmography ;Supervised Learning**



**doi:10.6342/NTU202403578**

# Contents





doi:10.6342/NTU202403578

doi:10.6342/NTU202403578

doi:10.6342/NTU202403578

# List of figures





doi:10.6342/NTU202403578

doi:10.6342/NTU202403578

# List of Tables





doi:10.6342/NTU202403578

# Chapter 1

# Introduction

Atrial fibrillation (AF) is an extremely common and clinically severe heart arrhythmia. The prevalence rate of AF is 0.5% among the worldwide population [46]. The macroscopic sign of AF involves the irregular and disorganised contraction of the atria, leading to compromised cardiac function and potentially impacting systemic blood circulation.

Stroke is a major cardiovascular incident with significant complications. Medical doctors diagnose stroke using various imaging methods, with Brain Computed Tomography (brain CT) being one of the most applied techniques for instant visualization. Stroke can be categorised into two types: ischaemic and haemorrhagic.

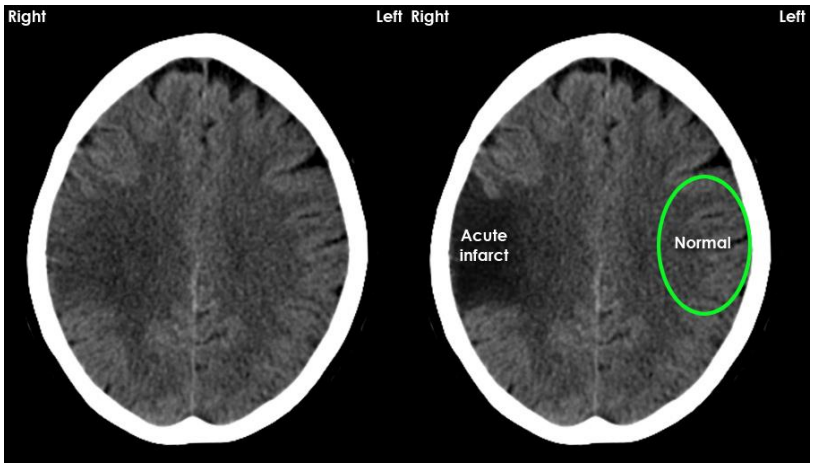
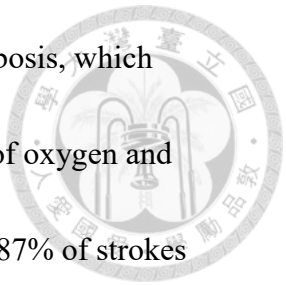



Figure 1.1 Schematic of brain CT of stroke patients reproduced from [49]



doi:10.6342/NTU202403578



An ischaemic stroke is triggered by clot aggregation or atherothrombosis, which narrows the blood vessels in the brain, impairing the transportation of oxygen and nutrients due to blood flow congestion. According to statistics [44], 87% of strokes are ischaemic. As shown in Figure 1.1, the impairment or even atrophy of blood vessels due to ichaemic stroke can be visualized by brain CT, providing valuable pathological information to medical professionals.

On the other hand, haemorrhagic stroke is generally more fatal. It occurs when a burst aneurysm or torn artery leads to intracerebral haemorrhage, causing bleeding in the cerebral tissues. This type of stroke can rapidly escalate in severity, often resulting in significant brain damage or death if not treated promptly.

AF is a common cardiac arrhythmia characterised by an irregular and often rapid heart rate. This condition arises when the heart's upper chambers, The atria, which receive erratic electrical signals, cause them to quiver or fibrillate instead of constricting properly. AF can lead to poor blood flow and is associated with a risk of serious complications, including heart failure and stroke. The erratic heartbeat may trigger symptoms as palpitations, shortness of breath, fatigue, and dizziness.



doi:10.6342/NTU202403578

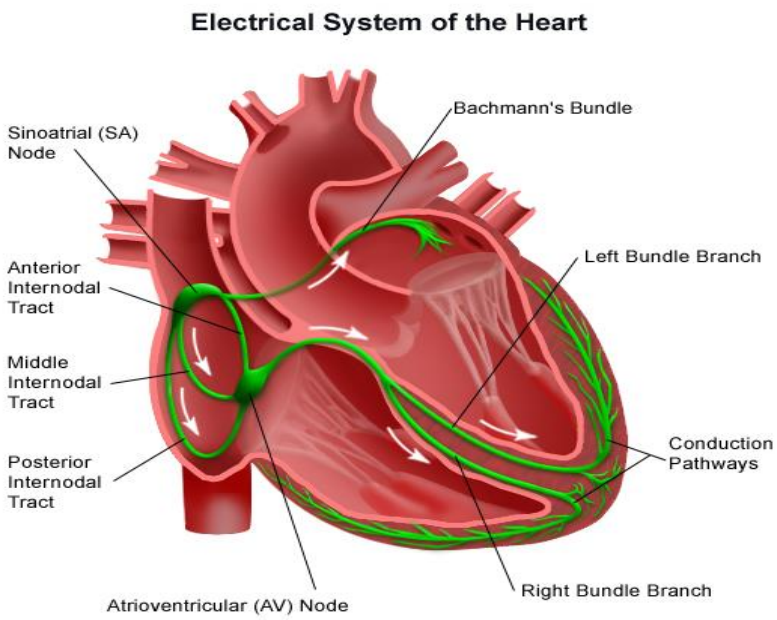



Figure. 1.2 Schematic of cardiac anatomy, reproduced from [34]

The Figure 1.2 illustrates the electrical conduction system of the heart, which is essential for maintaining a consistent and coordinated heartbeat. Functionally, the sinoatrial (SA) node, situated at the superior aspect of the right atrium, functioning as the primary pacemaker of the heart, generating electrical impulses that initiate atrial contraction. These impulses travel through the internodal tracts and Bachmann's bundle, which ensures simultaneous contraction of the atria.

Subsequently, the atrioventricular (AV) node, located near the interatrial septum, receives these impulses and acts as a gatekeeper, slowing them down before they reach the ventricles. This delay ensures that the ventricles have enough time to fill



doi:10.6342/NTU202403578

with blood from the atria before they contract. The bundle of His, along with the Purkinje fibbers, plays a crucial role in transmitting these electrical signals to the ventricular myocardium. The bundle of His splits carry the impulses to the both left and right ventricles. The Purkinje fibbers then distribute the electrical impulse throughout the ventricles, facilitating their synchronized contraction during the cardiac cycle.

Nonetheless, as people age, the electrical conduction systems of the heart gradually face challenges due to several potential issues. Some symptoms, such as asthma or compromised respiratory health, are indicative of inflammatory airway diseases. Asthma and Atrial Fibrillation (AF) share certain pathological similarities [36]. The inhalation and exhalation process involve the diaphragm and chest, facilitating the circulation of fluid (blood and gas) throughout the body. Clinically, the respiratory rate (RESP) refers to the number of breaths per minute, normally falling between 12 to 20 times. An outlier in the RESP value may serve as an indicator of a respiratory disease, necessitating consultation with medical professionals for proper evaluation and management.

To gain a deeper comprehension of impaired cerebral conditions, it is imperative to conduct a thorough analysis of blood pressure and flow rate, elucidating the



doi:10.6342/NTU202403578

complicated relationship between cerebral and cardiac functions. For reasons of convenience and cost-effectiveness, the standard practice entails the regular non-invasive blood pressure by stroke units in hospitals, a practice that is globally recognized and adhered to.

Tracking patients after thrombosis treatment for embolism is a standard approach to comprehensively understanding the progression of stroke over a three-month period. However, because of the challenges associated with determining the exact cause of death for patients who experience sudden death outside the hospital, as well as the potential for blank records due to non-return of patients and family members for follow-up appointments, this study aims to simplify its focus by excluding cases of stroke-in-evolution. Despite implications in certain references for example [3], this exclusion is intended to ensure a clearer and more manageable analysis of data related to atrial fibrillation (AF) and its associated outcomes.

In accordance with [13], the risk factors of AF include:

1.Progress of aging

2.Hypertension i.e. high blood pressure

3. Obesity



doi:10.6342/NTU202403578

4. Smoking

5. Hyperthyroidism

6. Chronic kidney disease

7. Other related heart diseases

Despite our limited understanding of the intrinsic metrics behind physiology, the literature suggests a strong correlation between sinus node dysfunction (SND) and atrial arrhythmia, as referenced in [13]. This correlation emphasises the need for more investigation to clarify the mechanisms of this relationship.

. It suggests that around 40%~70% SND are accompanied by atrial arrythmia.

## 1.1 Brain Structure and its Physiology

Despite its relatively small volume compared to the entire body, the brain is one of the most indispensable organs. It is liable for nearly all neurological responses, excluding those of the autonomic nervous system. The brain's compact size belies its intricate functions, which are of utmost significance for our bodies.

The brain includes millions of blood vessels, with the majority being capillaries.



doi:10.6342/NTU202403578

These capillaries play an important role in serving the inter-cellular metabolism, maintaining its proper physiology. Before blood reaches the capillaries, the diameter of blood vessels inevitably decreases as it transitions from larger arteries to smaller arteries. This process, known as vascular constriction, plays a critical role in regulating blood flow and distribution throughout the body.

There are four major divisions of the brain as Figure 1.3, Frontal, Temporal, Parietal and Occipital lobe. They serve different functions regarding neurological responses.

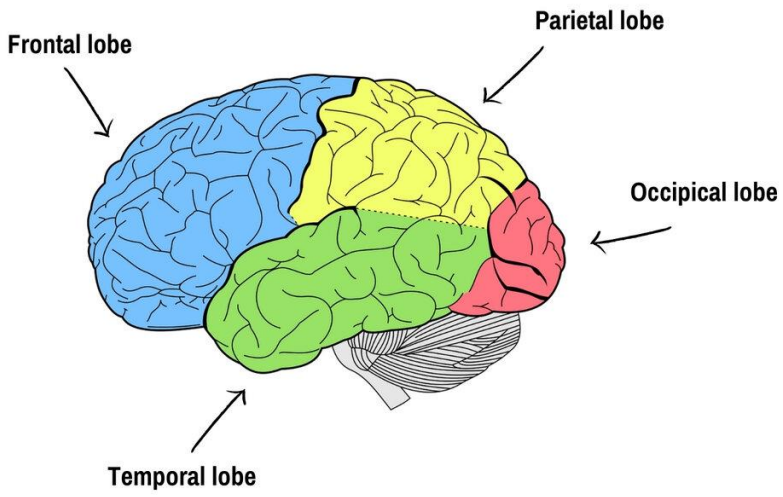



Figure 1.3 Schematic of cerebral lobes reproduced from [24]

The frontal lobe functions the problem-solving skills, memory and movement of the body serves as one of the indispensable parts of the brain, can certainly under threat if an ischaemic stroke attack.



doi:10.6342/NTU202403578

When it comes to the temporal lobe, it is responsible for emotion regulation, language processing, and memory storage, is significantly impacted by ischaemic stroke. Dysfunction in temporal lobe can induce emotional disabilities, impair language expression, and result in memory loss. These symptoms often serve as early indicators of dementia, a progressive condition marked by a decline in brain function and eventual physical disability leading to death.

On the other hand, the parietal lobe plays a crucial role in processing sensory stimuli, including touch, temperature, and spatial awareness. Its characteristics can be regard as the most instinctive on when compared to two lobes from the previous contexts.

Last but not the least part, the occipital lobe, is the processor of the visual information. It handles the sight received from eyes delivered by the optical nerves and makes necessary responses to adapt the surroundings.

## 1.2 Blood pressure

Systolic pressure is the pressure in the when the heart contracts to pump blood, representing the maximum pressure in a single cardiac cycle. In contrast, diastolic pressure is pressure in the arteries when the heart relaxes between beats, allowing the



doi:10.6342/NTU202403578

chambers to fill with blood, particularly deoxygenated blood returning from the body's circulation. Both two types blood pressures are vital indicators of cardiovascular well-being.

Consistent blood pressure within appropriate intervals is essential for ensuring a moderate blood flow rate, which, in turn, helps maintain the elasticity and structural integrity of blood vessels. Mean blood pressure (MBP) is defined as the sum of one third of SBP and two third of DBP is because of the approximate phase duration in single cardiac cycle. MBP is also essential for knowing the function of tissue perfusion.

$$\mathrm{MBP} = \frac{1}{3}\mathrm{SBP} + \frac{2}{3}\mathrm{DBP} \tag{1}$$

Another quantity is defined as Pulse pressure (PP). Pulse Pressure is the difference value between Systolic and Diastolic pressure, and in accordance with [32], there exists a positive correlation between PP and the cumulation of AF counts. As people get elder, the tendency of the PP is likely to increase because of the vessel stiffness induced higher SBP and lower DBP compare to early age.

$$PP = SBP - DBP \tag{2}$$



doi:10.6342/NTU202403578

## 1.3 Myocardial Infraction

Myocardial Infarction (MI) and AF share common risk factors. Research suggests that AF can result from atrial ischemia [18], indicating that understanding the ischemic response in the atria is crucial. This understanding is particularly important in the context of its potential impact on middle cerebral artery stroke as the major cause of ischaemic stroke, highlighting the significance of these interrelationships in an aging society.

Furthermore, it's important to acknowledge the relationship between Myocardial Infarction (MI) and AF. Both conditions represent substantial threats within the realm of cardiovascular disease and can precipitate undesirable outcomes. This underscores the necessity of comprehensive management strategies to address these interconnected health concerns effectively. As the [29] mentioned, the estimated prevalence of AF is approximately 0.60% in men and 0.37% in women globally. It is evident that men are more prone to developing AF.

## 1.4 Middle Cerebral Artery Stroke

Middle Cerebral Artery Stroke (MCA) is recognized as the most prevalent subtype of



doi:10.6342/NTU202403578

ischemic stroke, as documented in the literature [19]. This condition is characterised by the occlusion or obstruction of the cerebral blood flow within the larger artery, leading to abrupt impediment caused by a reduction in arterial diameter. As Figure 1.4, The figure shows the cerebral arteries forming a ring-like structure at the brain's base. This system ensures consistent blood flow to the brain, even if one artery is blocked.

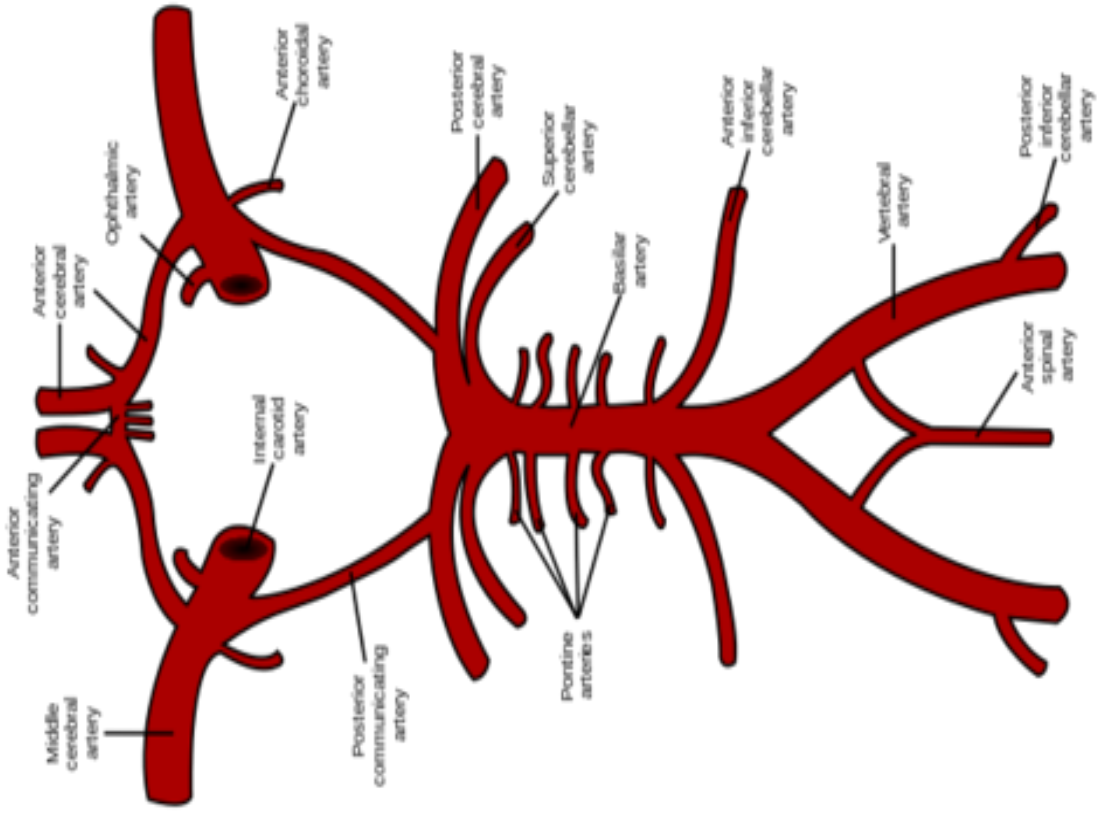



Figure 1.4 Schematic of arteries in the brain reproduced from [19]

Diameter changes in these arteries significantly impact brain health. Narrowing



doi:10.6342/NTU202403578

(stenosis) reduces blood flow, increasing ischaemic stroke risk. Conversely, widening (aneurysm) can lead to rupture and haemorrhagic stroke. This arterial arrangement offers a backup route for blood flow, maintaining brain perfusion under varying conditions. In brief, the blockage or clots stuck the MCA to impair the supplement of nutrients as well as oxygens.

Consequently, this event culminates in the impairment of neurological functions, manifesting as cognitive or physical disabilities, and in severe cases, mortality. The term 'ischaemic' pertains to the insufficient supply of oxygen to the brain tissues, instigating localized dysfunction within the cerebral region for example frontal, temporal and parietal lobes of the literature [23].

# 1.5 Modified Rankin Scale definition

The Modified Rankin Scale (mRS) serves as a globally recognised index for quantifying the neurologic disabilities by physical mobility and self-care abilities of individuals affected by stroke. It is customary for healthcare practitioners to reassess patients who have undergone thrombolysis treatment after a three-month interval, in order to evaluate their outcomes. However, it can be challenging to record for specific



doi:10.6342/NTU202403578

individuals who do not have follow-up appointment. Given that, we adopt the mRS value just after treatment is conducted for every single patient. The clinical physicians labelled out patients by straightforward observation.

The following are the indices with description of mRS and corresponding symptoms.

| mRS value | description |
|---|---|
| 0 | without any residual symptoms. |
| 1 | without significant disability, patients is able to carry out the routine as pre-stroke stage. |
| 2 | slightly disability, but patients can carry out most daily tasks without assistance. |
| 3 | moderate disability, external assistance of self-caring is needed but patients are still able to walk on his/her own. |
| 4 | moderate-severe disability, patients is not being able to walk without others' assistance. |
| 5 | severe disability which needs constant bed nursing care |



doi:10.6342/NTU202403578

| | |
|---|---|
| 6 | death |

Table 1.1 mRS description from [47]

There is another commonly-used index of National Institute of Health Stroke Scale (NIHSS), with domain between 0 to 42 which serves another dedicate prognosis in medical realm when it comes to stroke. However, the scoring is more laborious for clinicians and some of the records are not available in our case. However, this exhibits a strong correlation with mRS according to [43].

# 1.6 Risk Factors

## 1.6.1 Diabetes Mellitus

Diabetes Mellitus (DM) represents a significant cardiovascular concern due to its association with insulin resistance. Blood sugar refers the glucose in the blow flow playing a critical role of the circulatory system, serving as the primary energy source for cellular ATP production. Insulin resistance disrupts the regulation of blood sugar levels, impacting vascular health. In healthy individuals without DM, postprandial increases in blood sugar prompt pancreatic insulin secretion, which effectively lowers



doi:10.6342/NTU202403578

blood sugar levels. However, in insulin resistance, there is inadequate and less effective insulin secretion, leading to sustained hyperglycaemia and consequent inflammation within the vasculature of various organs, thereby accelerating the aging process.

Prolonged insulin resistance can progress to type II diabetes mellitus. While diabetes itself may not be inherently severe, it predisposes individuals to various health complications. For example, individuals with DM are at an increased risk of infection during wound healing and may experience sight loss due to Diabetic Retinopathy (DR). Effective healthcare management is crucial to mitigate these risks.

Moreover, DM significantly increases the risk of developing heart disease, including AF [21]. Drastic fluctuations in blood sugar levels and insulin resistance contribute to left ventricular hypertrophy, a pathological risk factor for AF development.

### 1.6.2 Hypertension

Hypertension (HT) is characterized as consistent blood pressure exceed proper range. According to the definition of World Health Organization, either the systolic pressure above 140 mmHg or diastolic pressure above 90 mmHg are regarded as HT. HT is



doi:10.6342/NTU202403578

also a common and easily measured factor when it comes to risk factors to all cardiovascular disease.

Pathological symptoms include the Atherosclerosis which means the narrowing of lumen inside blood vessel. There is another type of HT is triggered by wall thickening. The wall thickening is

an increase in the diameter of vessel walls can pose a significant threat to cardiovascular health. Both congenital and acquired wall thickening contribute to elevated blood pressure. This stiffness in vessel walls forces the heart to pump more vigorously to maintain blood flow, leading to increased strain on the heart over time.

Dynamic cerebral autoregulation(dCA) is the ability of cerebral vessels regulate themselves in response to the systemic blood pressure fluctuation. However, in individuals with chronic HT, the range within which dCA operates can shift. Specifically, chronic hypertension can lead to a reconfiguration of the autoregulatory range, such that the cerebral vessels become adapted to higher blood pressure levels for maintaining stable cerebral perfusion

Chronic issues triggered by HT such as mitral stenosis and arrhythmia can result from compromised heart health. As cited in [42], studies confirm a correlation between AF



doi:10.6342/NTU202403578

and left vessel wall thickness.

## 1.6.3 Atherosclerosis

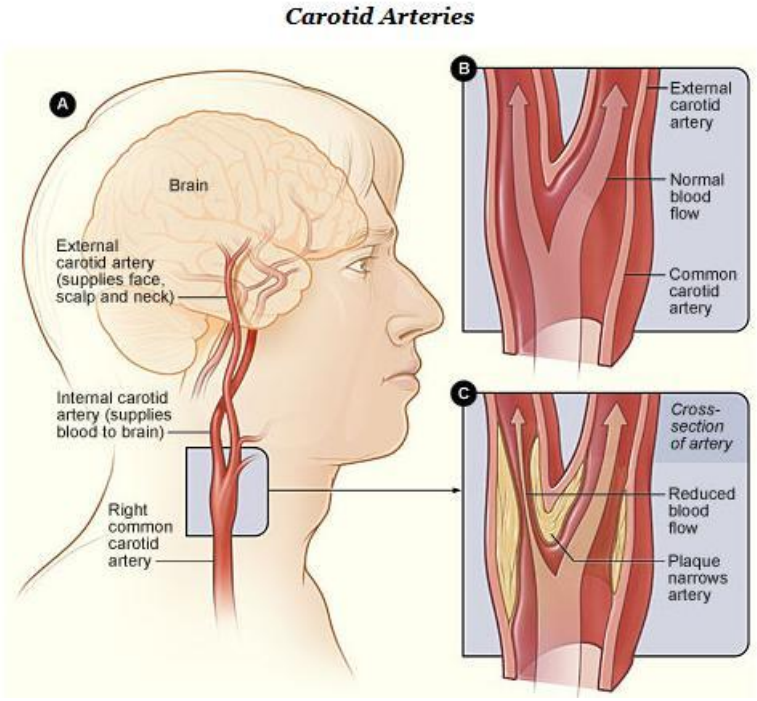



Figure 1.5 Schematic diagram of Atherosclerosis reproduced from [48]

Atherosclerosis represents a significant risk factor for the progression of AF. It is characterized by the gradual accumulation of plaque, known as Atheroma, within the arteries. This process leads to the hardening of endothelial cells, impairing the vessel wall's autoregulation function and increasing overall vessel stiffness. The plaque's composition, often rich in low-density lipoprotein (LDL) and other fats, contributes to



doi:10.6342/NTU202403578

the progression of atherosclerosis. Hyperlipidaemia, characterized by elevated levels of LDL, is a key pathological feature associated with atherosclerosis, often resulting from the overconsumption of animal fats and poor

weight management. As shown in Figure 1.3, the schematic diagram compares a normal situation with plaque accumulation near the neck, which can obstruct the carotid artery. As mentioned in [48], air pollution can increase the risk of atherosclerosis.

Maintaining a healthy lifestyle, including regular exercise and proper weight control, is essential for preventing hyperlipidaemia and reducing the risk of atherosclerosis. Furthermore, atherosclerosis poses a significant threat to coronary artery disease, which needs early prevention.

# 1.7 Classification

In medical studies, sensitivity refers to the capability of a test to correctly identify those with the disease or condition. It measures how well the test correctly identifies true positives, avoiding false negatives. In the contrast,specificity, is the capability of a test to accurately identify those without the disease or condition. It measures how



doi:10.6342/NTU202403578

well the test correctly identifies true negatives, avoiding false positives. Both sensitivity and specificity are important because they indicate how well a test can accurately diagnose the presence or absence of a disease, which is crucial for effective medical decision-making.

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

The upper two equations are the definition of sensitivity and specificity of binary classification as [1] did. For our interests, we will expand our paragraph to make effort to further classify into 3 truncated group based on the Modified Rankin Scale (mRS) definition in our task making an advancement based on the contribution of [1].

The Receiver Operating Characteristic (ROC) curve serves as a quantitative measure for assessing the performance of a classifier, offering a nuanced evaluation beyond simple accuracy calculations. The Area under the ROC curve (AuROC) represents the extent to which the classifier compromises between sensitivity and specificity in a classification task. An AuROC value of 0.5 signifies a classifier that performs no better than random chance, akin to tossing a coin as the diagonal dash line of Figure 1.5. Conversely, an AuROC of 1.0 denotes a perfect classifier that makes all



doi:10.6342/NTU202403578

predictions correctly, without any false positives or false negatives.

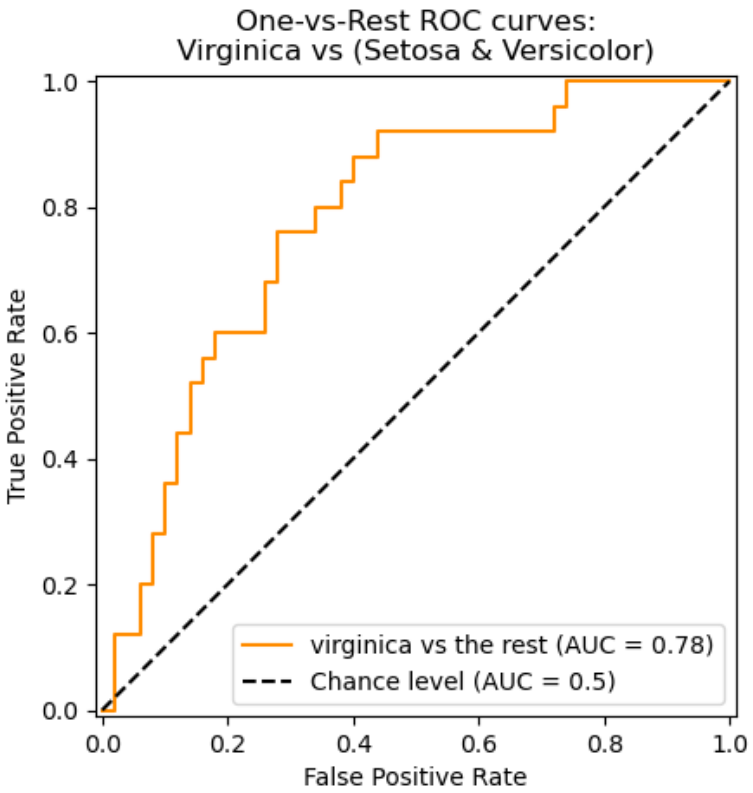



Figure 1.5 Schematic of ROC curve reproduced from [50]

In practice, the AuROC value typically falls between these two extremes, reflecting the classifier's performance relative to random chance and ideal classification. Functionally, the AuROC value demonstrates the model's capability to differentiate between different classes and determine the trade-off between true positive and false positive rates.



doi:10.6342/NTU202403578

# 1.8 Machine Learning and Deep Learning

Machine Learning (ML) and Deep Learning (DL) are both popular AI techniques to solve classification problems in various fields. In ML, features are typically crafted by humans based on domain knowledge. These features are then used as inputs to the model, which learns the mapping between these features and the output labels. This approach requires domain expertise to design relevant features. On the other hand, in DL, the model automatically learns features from the raw data. Deep neural networks are capable of automatically discovering intricate patterns exploring features from the data, removing the need for manual feature engineering. This is one of the key advantages of DL, as it can potentially lead to more effective models without the need for domain-specific feature design.

A common mathematical type of DL is called Convolutional Neural Network (CNN),

CNN can be a powerful way to computation which involves an input layer, designated numbers of hidden layers and an output layer for making decision especially for visual pattern or higher dimensional data.

The limitation posed by the small sample size necessitates the use of ML over DL in our study. DL's efficacy is contingent upon substantial amounts of data for effective



doi:10.6342/NTU202403578

and accurate learning. Moreover, DL models, operating on extensive datasets, often yield metrics that are intricate and challenging to fully interpret, potentially leading to less intuitive insights. Given the relatively small nature of our dataset as well as the existence of artifacts, the feasibility of employing DL is deemed poor. Thus, the decision to utilize ML aligns with the pragmatic need to work within the constraints of limited data availability.

## 1.9 Conclusions

The deleterious effects of acute strokes on both short-term and long-term physiological responses underscore the need for a quantitative approach to establish key indices and their interrelationships. As previously discussed, these risk factors are under scrutiny to elucidate their correlations and warrant further investigation to advance medical understanding.

This emphasis on quantitative analysis serves to enhance our awareness of the intricate interplay between various risk factors, offering insights that can potentially inform clinical practice and therapeutic interventions. Thus, further studies are imperative to substantiate these findings and propel medical advancements in stroke



doi:10.6342/NTU202403578

management.

We will break down the features engineering step by step in the next chapter for instance filtering and thresholding. Further, we discuss how it works after statistical method is applied and the evaluation of the performance and inspiration in the end.



doi:10.6342/NTU202403578

# Chapter 2

# Materials and Methods

In this Chapter, we introduce some measurements and implement our proposed analysis computationally. We also combine the domain knowledge of medical professionals and ML algorithms to investigate the significance of certain features and optimize the prediction outcome regarding to the patient with pathological symptoms, as introduced in Chapter 1.

## 2.1 Data acquisition

The physiological data are measured by the IntelliVue Bedside patient monitor by Philips, which includes Electrocardiography (ECG), Plethysmography (PPG), and NBP (Non-invasive blood pressure) measurements with sampling rates of 512 Hz, 128 Hz, and 128 Hz, respectively. The ECG is measured using standard 12-lead equipment widely recognized in cardiology departments around the world.

## 2.2 Population Demographics



doi:10.6342/NTU202403578

Initially, the dataset included a comprehensive cohort of 221 patients with both ECG and PPG records. However, due to absent records of continuous NBP monitoring for some patients, 109 instances were excluded. This resulted in a refined subset comprising 112 patients, ensuring a more focused and optimised dataset for analysis.

| mRS | 0~2 | | 3~4 | | 5~6 | |
|---|---|---|---|---|---|---|
| gender | men (n =6) | women (n = 5) | men (n =23) | women (n=20) | men (n=30) | women (n=28) |
| age (mean±SD) | 54.33± 12.94 | 74.60± 5.68 | 71.43± 10.95 | 72.85± 11.95 | 74.73± 12.08 | 79.79± 8.33 |
| SBP (mmHg) (mean±SD) | 147.83± 12.94 | 157± 5.68 | 161.74± 10.95 | 173.33± 11.95 | 156.57± 12.08 | 167.85± 8.33 |
| DBP (mmHg) (mean±SD) | 83.33± 8.59 | 85± 23.54 | 90.30± 19.27 | 89.61± 25.21 | 86.32± 15.63 | 87.22± 20.08 |
| HR (beats/min) (mean±SD) | 98.67± 27.37 | 88.4± 37.37 | 100.04± 25.82 | 82.65± 23.06 | 83.52± 25.90 | 82.21± 16.79 |
| Hypertension % | 66.67 | 80 | 82.61 | 85 | 86.67 | 85.71 |
| Diabetes % | 33.33 | 0 | 39.13 | 40 | 36.67 | 17.86 |

Table 2.1 Demographics of the patients



doi:10.6342/NTU202403578

| | p-value of mean age | p-value of mean SBP | p-value of mean DBP | p-value of mean HR | p-value of mean HT% | p-value of mean DM% |
|---|---|---|---|---|---|---|
| gender | 0.3219 | 0.7390 | 0.6792 | 0.6006 | 0.4297 | 0.7133 |
| Severity | 0.2326 | 0.3135 | 0.6336 | 0.0858 | 0.1785 | 0.7032 |

Table 2.2 table of two-way ANOVA p-values among patients

Table 2.1 is the statistics of our patients in and their physiological parameters from ICU. We also list out both DM and HT percentage for different gender to display the characteristic of our patients. On the other hand, Table 2.2 presents the results after multiple comparisons of our data to identify potential correlations based on our classes. In our analysis, there are 12 tests (as the same groups of Figure 2.1) and the corresponding p-values were calculated using two-way ANOVA. Two-way ANOVA is a method used to assess correlations between physiological parameters and our classes based on gender and truncated stroke severity. We calculated p-values by comparing the variation between groups (sum of squares) to the variation within groups (error sum of squares), with a significance level set at 0.05 to determine statistical significance.



doi:10.6342/NTU202403578

This indicates that no physiological parameter showed statistically significant correlations with age and truncated stroke severity in our analysis. This outcome suggests that we need to investigate more complex physiological factors rather than relying solely on the few mentioned physiological parameters.

## 2.3 Electrocardiography signals

Electrocardiography (ECG) measurement records the electric potential of the heart using electrodes placed on the skin. It captures the heart's rhythm, rate, and electrical conduction patterns, producing a waveform that reveals the timing of atrial and ventricular contractions. Clinically, ECG is crucial for diagnosing arrhythmias, myocardial infarction, and other cardiac abnormalities. It aids in monitoring heart conditions for clinical doctors. By analysing ECG data, healthcare providers can detect heart disease early, prevent complications, and improve patient outcomes.



doi:10.6342/NTU202403578

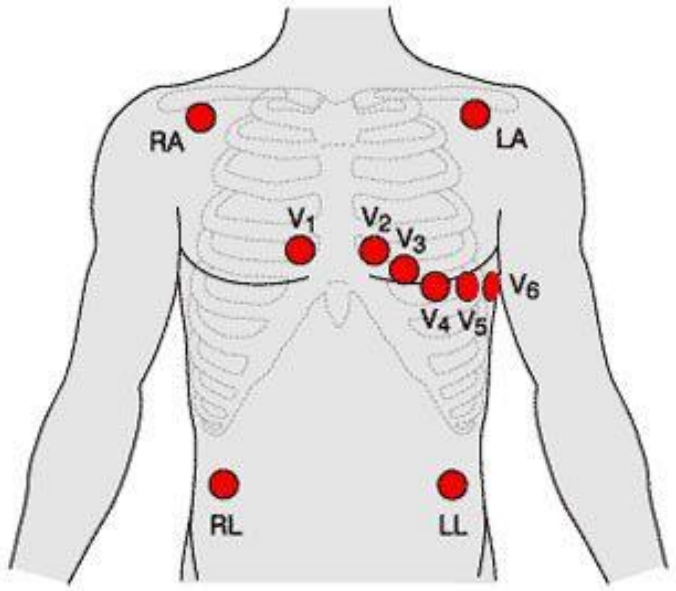



Figure. 2.1 Schematic diagram of 12 lead ECG measurement reproduced from [25]

The upper plot demonstrates how the electrode are placed on patients. The 12-leads means that there are 12 perspectives to monitor the heart's activity.

The elaboration where each lead is positioned is as follows:

V1: Placed in the fourth intercostal space just to the right of the sternum. It provides a view of the heart's electrical activity from a right-sided perspective.

V2: Placed in the fourth intercostal space just to the left of the sternum. It provides a view similar to V1 but from a slightly more left-sided perspective.

V3: Placed between V2 and V4, halfway between the positions of V2 and V4. It



doi:10.6342/NTU202403578

provides a view that is balanced between right and left perspectives.

V4: Placed in the fifth intercostal space at the midclavicular line (a vertical line that runs through the midpoint of the collarbone). It provides a view of the heart's electrical activity from a direct frontal perspective.

V5: Placed horizontally in line with V4 at the anterior axillary line (a vertical line that runs through the front of the armpit). It provides a view of the heart's electrical activity from a left anterior perspective.

V6: Placed in line with V5 but at the midaxillary line, are used to assess different regions of the heart's left ventricle, as well as the main pumping chamber. They help in diagnosing various heart conditions, such as myocardial infarction (heart attack), ischemia and arrythmia etc.

Heart Rate Variability (HRV) refers to the variations between successive heartbeats. There are several statistical methods that can reflect HRV, which we will break down in the subsequent paragraphs. It is a key focus in the analysis of Electrocardiogram (ECG) signals due to its significance in investigating the autonomic nervous system. The autonomic nervous system is regulated by two major divisions: the sympathetic and parasympathetic nervous systems, which are antagonistic to each other. The



doi:10.6342/NTU202403578

sympathetic nervous system promotes cardiac pulsation, whereas the parasympathetic nervous system calms it.

A considerable amount of research concludes that a higher HRV is a more desirable outcome for patients since a higher HRV indicates the heart regulate itself to adapt external stimulation or different physical activity status. Nonetheless, the study [12] indicates that certain pathological symptoms can also trigger a higher HRV also our interests of patients under arrhythmia (specifically AF) condition which means that proper HRV is preferred cases for a healthier condition we wish to see.

### 2.3.1 Pan-Tompkins filtering

The Pan-Tompkins Algorithm is a well-established method for ECG signal processing, widely utilized for the precise identification and extraction of the characteristic PQRST waves within each cardiac cycle. The standard Pan-Tompkins filtering methodology comprises a sequential application of several key stages: a low-pass filter to remove high-frequency noise, a high-pass filter to eliminate baseline wander, differentiation to accentuate the QRS complex, squaring to enhance the true signal we desire for, and moving window integration (MWI) to amplify the QRS



doi:10.6342/NTU202403578

complex. The final step traditionally involves adaptive thresholding to detect R-peaks accurately.

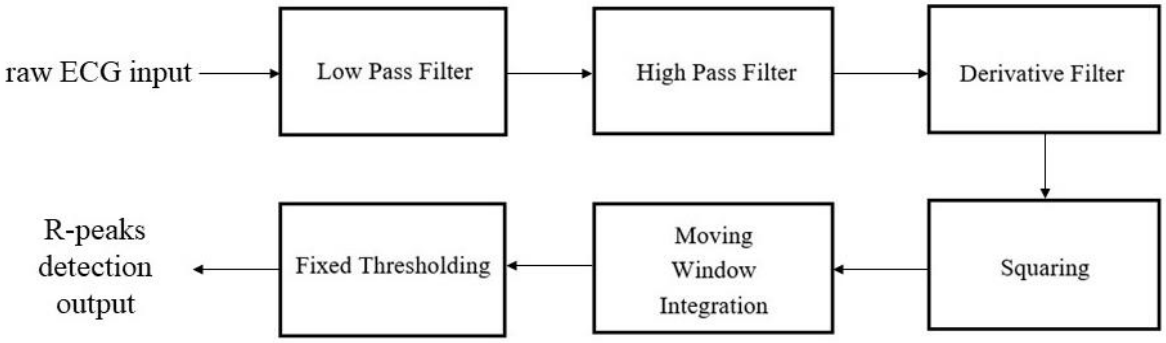



Figure 2.2 flowchart of Simplified Pan-Tompkins Filtering

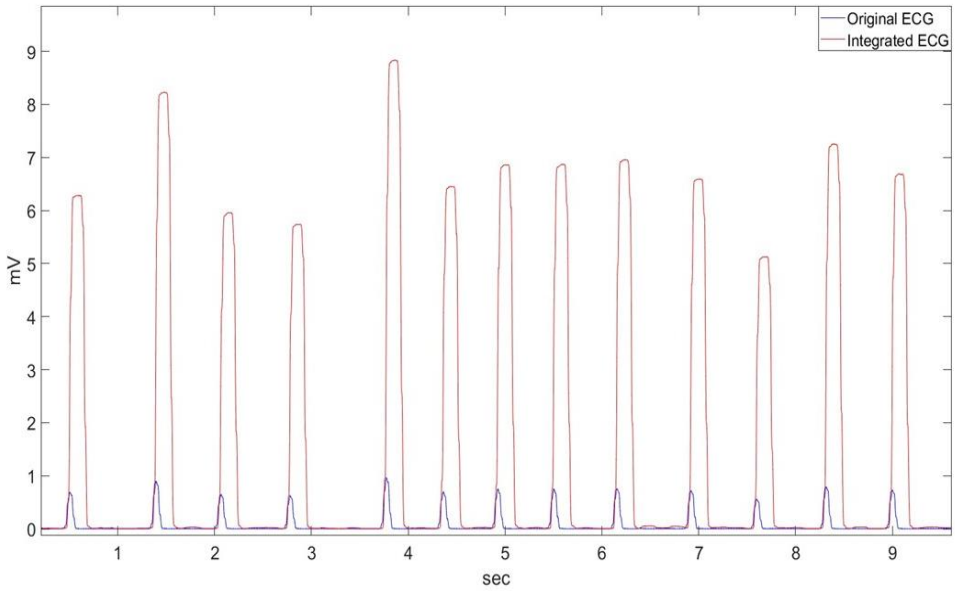



Figure 2.3 Schematic of Moving Window Integration



doi:10.6342/NTU202403578

As shown by the upper schematic diagram, the approach taken involves filtering and denoising ECG signals with the aim of accentuating the QRS complex while reducing the prominence of other waveforms, such as the P and T waves. This strategy is particularly pertinent for enhancing the visibility of patterns in the PQ and ST segments, which are known to manifest irregularities in patients with AF. The goal is to improve the efficacy of feature extraction by mitigating noise from other phases of the cardiac cycle.

In our study, we have specifically incorporated a fixed area under the curve threshold following the moving window integration step of the Pan-Tompkins Algorithm. This thresholding approach allows for the straightforward identification of R-peaks in ECG signals, enhancing the algorithm's efficiency and applicability. Further, for the purpose of the proper sampling the data, we also take median values after 5 blocks for each 5 minutes length of every single patient.

## 2.3.2 Time domain analysis

To address the issues about time domain analysis especially ECG, cardiac electrophysiology is key to understanding how the heart's electrical signals coordinate



doi:10.6342/NTU202403578

its contractions. This process, from the SA node initiating the signal to the coordinated contraction of the atria and ventricles, is critical for heart function. Analysing ECG signals helps detect abnormalities like arrhythmias, providing crucial insights into heart health.

**P** is generated by the depolarization of the atria, signify the initiation of atrial contraction. This contraction facilitates the transfer of blood from the atria to the left ventricle, a critical step in preparing to distribute oxygenated blood throughout the body's circulation.

**QRS** complex is the duration time of ventricular depolarization which triggers the contraction of ventricle and repolarization of the atria recovering from the contraction. QRS complex is the systolic period of heart systole, a steeper waveform i.e. a shorter duration of QRS complex can be a healthy indicator in cardiology suggested by [41].

**T** is the stage of ventricular repolarization which means the deoxygenated blood of systemic circulation will return to the atria and complete the cardiac cycle.

The discrimination between healthy people and AF patients often relies on the clarity of PQRST points, in contrast, AF characteristics can be the fluctuation of PQ and ST segments which means the heart cannot efficiently dispense and accommodate the



doi:10.6342/NTU202403578

blood flow respectively.

The original feature for instance RR-Intervals, is one of the most straightforward features for calculation. Its nature is the reciprocal of the heart rate per seconds, given by the time difference between adjacent R waves, shown in Figure 2.4. It is very intuitive for RR-Intervals extraction and visually simple just to find every local maximum of each cardiac period.

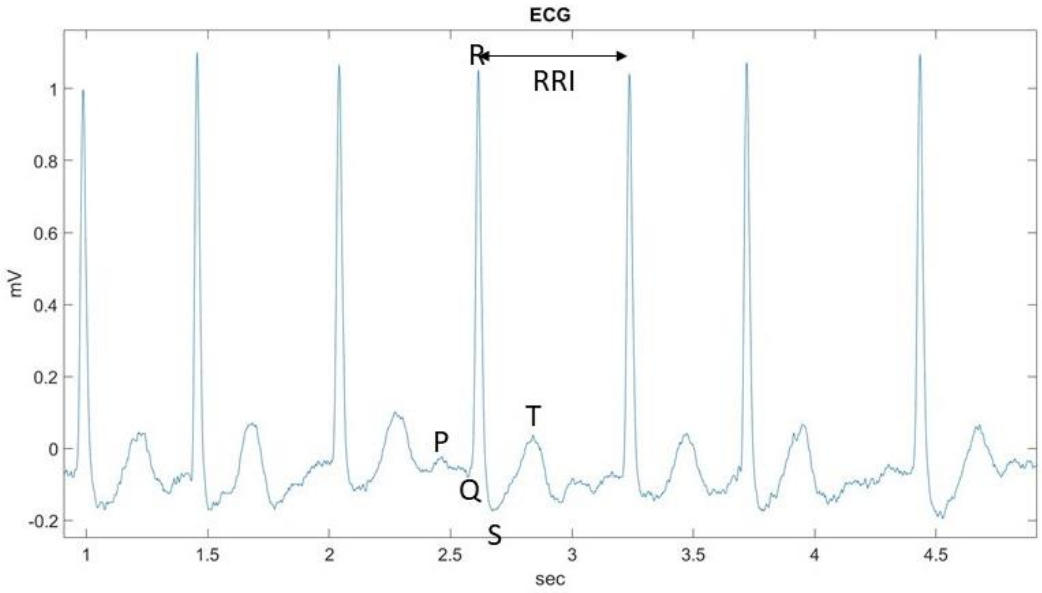



Figure 2.4 Schematic of phase of cardiac cycle

In addition to deriving various features from RR-intervals in a statistical manner, the Heart Rate per minute (HR) still serves as a pivotal quantitative metric for validation purposes. Its intuitive nature and quantitative representation make it an indispensable



doi:10.6342/NTU202403578

aspect in ensuring the accuracy and reliability of the derived statistical features from RR-Intervals.

The following features are derived from the RR-Intervals as Figure 2.4 shows.

| | Mathematics |
|---|---|
| SDRR | standard deviation of the RR-Intervals |
| pNN50 | the portion of successive RR-Intervals larger than 50 milliseconds |
| Root Mean Square Successive Difference (RMSSD) of successive RR-Intervals | $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x_{i-1})^2}$ where x is substituted by the series of RR-Intervals |

Table 2.3 HRV indicators in time domain

SDRR and RMSSD are indicative of short-term and long-term variability of heartbeats, respectively. pNN50, on the other hand, reflects the ratio of RR-Intervals differing by more than 50 milliseconds, providing a linear parameter for heart rate variability. Those parameters are mentioned by [12].



doi:10.6342/NTU202403578

## 2.3.3 Frequency domain analysis

Frequency Domain features are derived from the Power Spectrum Density (PSD), obtained through the Fast Fourier Transform (FFT), which quantifies the distribution of power within the frequency components of the signal. These features are often utilized to assess the autonomic nervous system's activity.

| | Frequency interval (Hz) | Physiological meaning |
|---|---|---|
| Very Low Frequency | 0.0003~0.04 | unknown |
| Low Frequency | 0.04~0.15 | sympathetic activity |
| High Frequency | 0.15~0.4 | parasympathetic activity |
| LFHF ratio | | sympathicovagal balance |

Table 2.4 Frequency domain parameters and physiological meaning

However, in our study the Frequency domain features of our patients are too similar which makes it less valuable to be in the list of features. For compromise, we excluded all the features from Frequency domain despite its clinical implication from other major studies like [2].



doi:10.6342/NTU202403578

## 2.3.4 Entropy domain analysis

Entropy Domain features are stem from further calculation of the time series features.

Entropy domain features in ECG and blood pressure analysis are valuable for assessing complexity and irregularity in physiological signals. They quantify the degree of disorder or unpredictability, offering a different perspective from traditional amplitude or frequency-based metrics. This similarity underscores their importance in capturing subtle changes and abnormalities that may not be evident through other methods, thus enhancing the diagnostic and prognostic capabilities of medical analyses.

**Shannon Entropy**

Shannon Entropy is a brief and intuitive approach to quantify the randomness or to say uncertainty of the region of interested data, the higher values of Shannon Entropy manifests a higher complexity patterns with variation of time series., where $\boldsymbol{x}$ represents the event symbol of the data series, such as the RRI (RR-interval) series extracted from the ECG signals. The data series length is represented by **n**. The term



doi:10.6342/NTU202403578

*H(x)* stands for the Shannon Entropy, and $P(x_i)$ is the probability function of occurrence of the ***i***-th event or symbol in the data set.

$$H(x_i) = \Sigma_{i=1}^{n} P(x_i)\, log_2(x_i) \quad (5)$$

**Multi-scale Sample Entropy analysis on ECG and ABP**

Multi-scale sample entropy (MSSE) is a common approach shown as [1] to quantify the problem dealing with biostatistics particularly in medical research. We adopt the conventional coarse graining approach to time scales from 1 to 20 scales for the non-invasive blood pressure to investigate the relationships between MSE and the stroke severity in mRS standard.

We quote the definition from reference [3], where the complexity is defined as the area under the curve of MSSE plot. According to the result provided, it indicates that the higher complexity of RR-Intervals and both SBP and DBP, the better the patient outcome.

The coarse grained $y_j$ can be written as the follows:

$$y_j^{\tau} = \frac{1}{\tau}\left(\sum_{i=(j-1)\tau+1}^{j\tau} s_i\right), 1 \le j \le \frac{N}{\tau} \quad (6)$$



doi:10.6342/NTU202403578

$y_j^\tau$: The $j$-th value of the coarse-grained time series at scale $\tau$. This value represents the average of $\tau\tau$ consecutive values of the original time series.

- $\tau$: The scale factor, which determines the number of consecutive values of the original time series that are averaged to calculate each value of the coarse-grained time series. In our case, set $\tau$ equals to 20, meaning each value of $y_j^\tau$ is the average of 20 consecutive values of the original time series.
- $N$: The length of the original time series. In your case, this represents the total number of data points in the NBP signal.
- $s_i$: The $i$-th value of the original time series. These are the individual data points of the NBP signal.

After the Multi-scale Sample Entropy calculation of both systolic and diastolic pressure, we conducted an unpaired t-tests with the Bonferroni-Holm Correction[1]. We

[1] The Bonferroni-Holm correction is a method to reduce false positives when conducting multiple statistical tests. It adjusts the significance threshold for each test by dividing the desired significance level the threshold α by the number of tests. This helps maintain the family-wise error rate (FWER).



doi:10.6342/NTU202403578

set significance level α =0.05. The analysis was visualised through error bar plots, which displayed the mean MSSE values for both systolic and diastolic measures. The two plots are Figure 2.5 and Figure 2.6 manifest downward trend from scale first to 20$^{th}$ scale. The ungrouped t-test result shows the first scales are both statistically significant by the threshold Bonferroni-Holm corrected p-values equals to 0.0028.

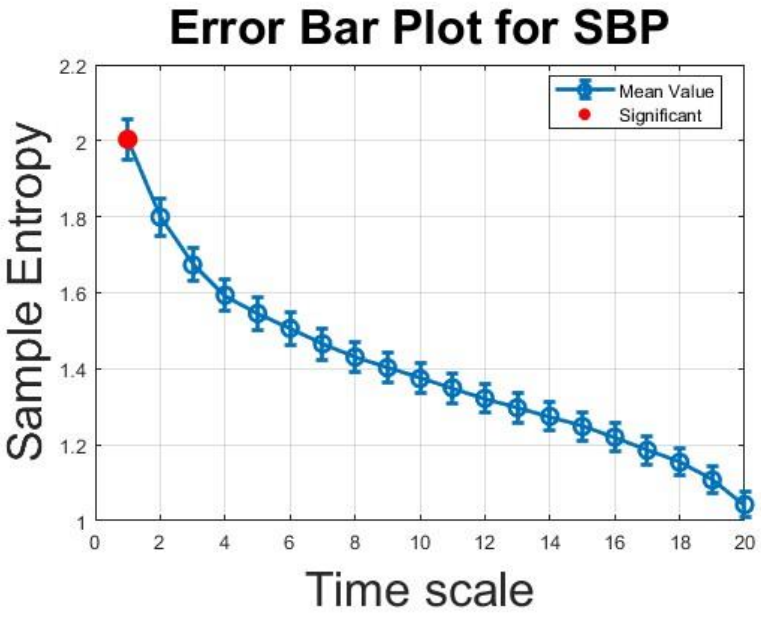



Figure 2.5 Error bar plot of systolic pressure



doi:10.6342/NTU202403578

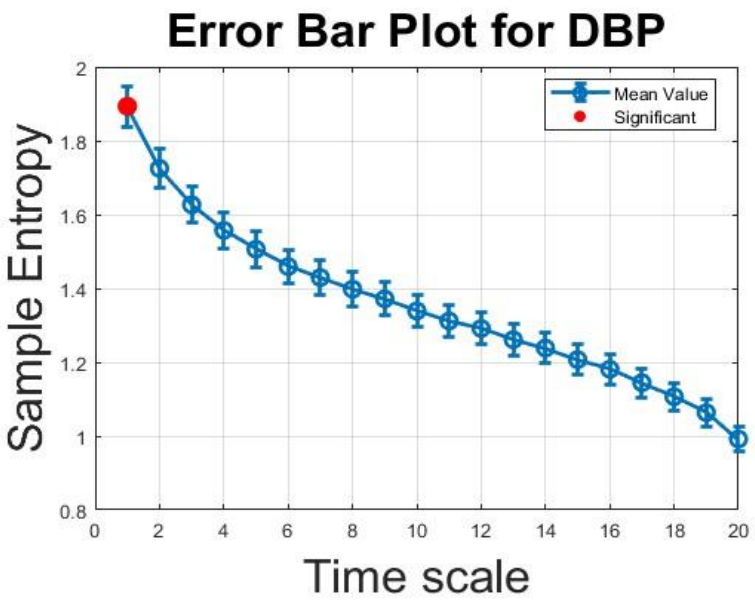



Figure 2.6 Error bar plot of diastolic pressure

## 2.4 Photoplethysmography signals

Photoplethysmography (PPG) stands as a pivotal non-invasive optical methodology widely employed in the realm of biomedical instrumentation for the assessment of various cardiovascular parameters. The clarity of measurements with minimal noise allows us to easily navigate physiological situations without the need for extensive filtering. This technique operates on the principle of photoplethysmography transduction, where in alterations in blood volume within the microvascular bed of tissues are elucidated through the modulation of light transmission characteristics. Such modulation arises from the differential absorption and scattering properties of



doi:10.6342/NTU202403578

biological tissues in response to incident light, thereby offering insights into physiological phenomena such as cardiac pulsations and peripheral perfusion dynamics. In practice, PPG entails the placement of optoelectronic sensors, typically in close contact with the dermal surface, most commonly upon the fingertips of subjects as the Figure 2.7.

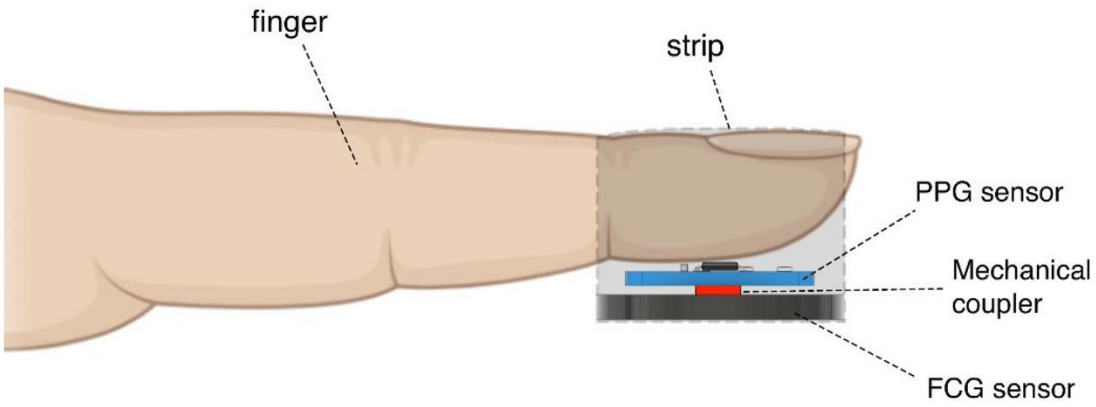



Figure 2.7 Schematic diagram of measurement of PPG reproduced from [45]

In the context of our research, the photoplethysmography (PPG) measurement offers an avenue to glean insights into Pulse Oximetry (PO), a valuable metric for assessing oxygen saturation in the blood. However, our study's primary focus lies elsewhere, as the direct relationship between variations in PO values and stroke outcomes presents a significant challenge. Therefore, while PPG may indirectly provide PO data, our research deems this aspect less pertinent, as the nuances in PO values may not be directly indicative of the stroke outcome dynamics we seek to investigate.



doi:10.6342/NTU202403578

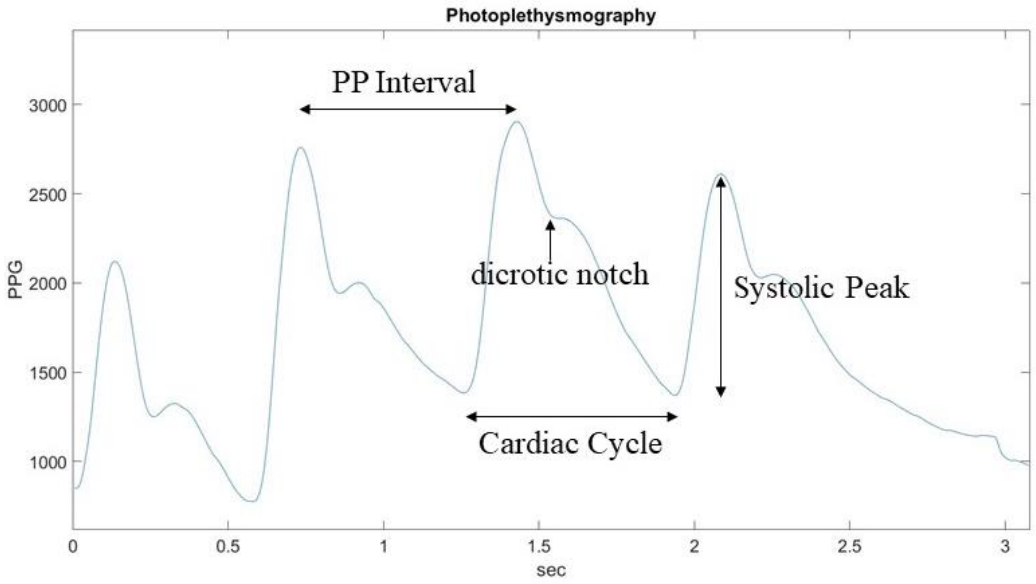



Figure 2.8 Schematic diagram of Photoplethysmography

As Figure 2.8 shown, the pattern of PPG is alike to the ECG, the peaks of PPG indicate the systolic movement and vice versa the valleys indicate the diastolic phase of the heart which squeezing the blow flow into the arteries initiating the process of carrying oxygens and nutrients to cells.

As the metrics involved between three measurements of our study, we also adopt the interception between the PPG and ECG to gain the Pulse Transit Time (PTT) to elaborate the analysis for a further depth of investigation.

PTT is the time difference between the heart squeezing the oxygenated blood to peripheral arteries can serve as an indicator estimating the stiffness of blood vessels



doi:10.6342/NTU202403578

the inverse idea against vessel compliance according to the reference [14].

Two types of the PTT are contained, the Peak to Peak as well as the PTTp is the well-recognized and most intuitive to have a brief insight.

On the other hands the Peak to the local maxima slope is named as the PTTs.

PTTs is the time difference between the R-wave and the maximum slope of PPG wave. PTT is an approach to quantify the vessel compliance for various studies including [39]. We made a graphical schematic diagram as Figure 2.9 to manifest the both the PTTp and PTTs.

Vessel compliance refers to the capability of blood vessels to adapt to dynamic changes in blood pressure and flow rate. The autoregulation of blood is essential of vessel compliance functions. The cohesive interaction between the endothelial layer and adjacent muscle tissues (smooth muscles) surrounding blood vessels plays a vital role in facilitating autoregulation. A study cited as [40] suggests that left atrial compliance can provide insights into the outcomes of AF and related cardiovascular diseases.



doi:10.6342/NTU202403578

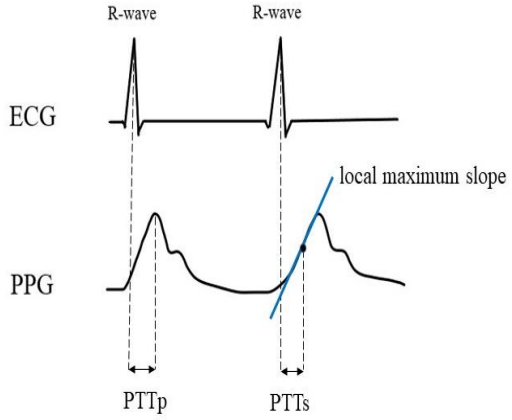



Figure 2.9 Schematic diagram of Pulse Transit Time

As we can see, the PTTp is clearly greater than the PTTs within a single cardiac cycle from the Figure 2.9. Also, we also applied certain statistic approach to describe the distribution of them. We adopted mean values and SD of both PTTp and PTTs as the features of our ML tasks to include their distribution.

## 2.5 Non-invasive Blood Pressure Analysis

Conventionally, Atrial Blood Pressure (ABP), particularly in its non-invasive form, is widely utilized as a primary measurement for simple cardiovascular disease prevention. The benefits of NBP measurements are cost-effective, convenience and high-fidelity immediate monitoring. The blood pressure, a fundamental macroscopic



doi:10.6342/NTU202403578

physiological parameter, plays an indispensable role in maintaining overall well-being. SBP denotes the peak arterial pressure during cardiac contraction, whereas DBP signifies the valley pressure between heartbeats. Research suggests that both SBP and DBP are vital indicators of cardiovascular health. Notably, studies have shown that SBP may serve as a more sensitive marker for cardiovascular disease compared to DBP [20]. Additionally, abnormal fluctuations in blood pressure, whether low or high, can manifest as potentially severe symptoms such as dizziness or even shock, underscoring the critical importance of maintaining blood pressure within a healthy range.

However, researchers and clinicians are increasingly exploring the hidden information within ABP waveforms. This information could serve as an indicator for investigating potential diseases associated with specific symptoms in clinical studies. Such efforts aim to better bridge the relationship between ABP waveforms and both the chronic degeneration of the circulatory system function or acute cardiovascular events such as stroke and heart attack.

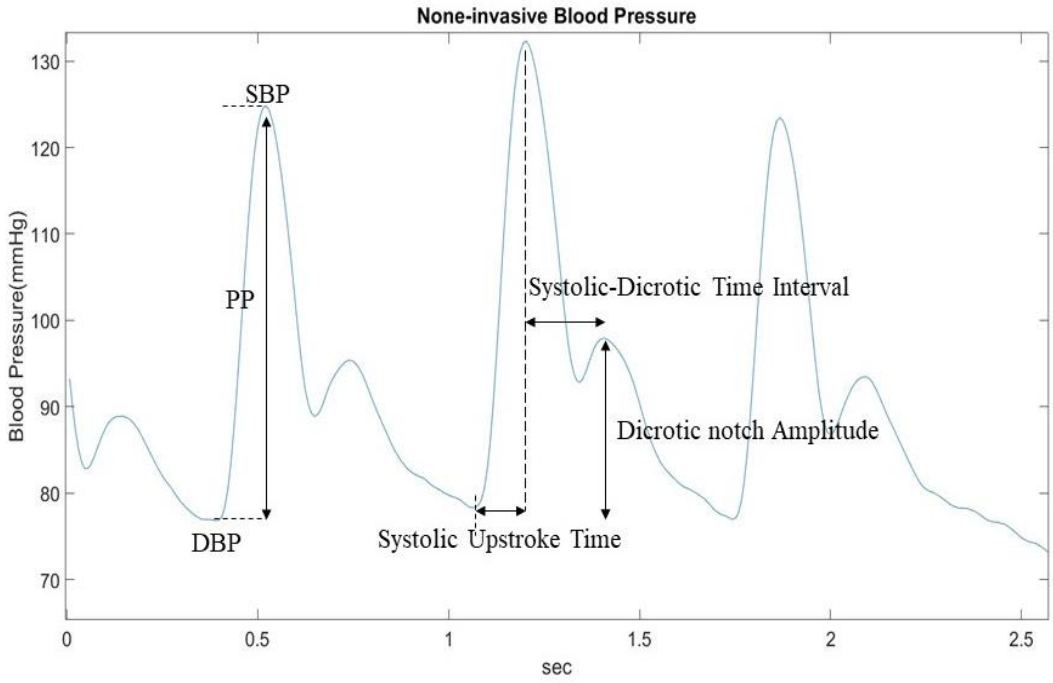
We adopted the concept from a previous study [8] which examines the variation in upstroke time as a potential physiological indicator for identifying the likelihood of coronary artery disease. His findings suggest that a shorter upstroke time is indicative



doi:10.6342/NTU202403578

of a healthier coronary condition in patients.

Systolic upstroke time (SUT), or as UT, is the temporal interval between the initiation of arterial blood pressure and the systolic upstroke, which corresponds to the peak pressure within a single cardiac cycle. This metric reflects the duration required for the pressure wave to propagate from the heart to the peripheral arteries. This parameter is influenced by various physiological factors, including arterial compliance, vascular resistance, and cardiac contractility, and it serves as a valuable indicator of cardiovascular dynamics and health.






doi:10.6342/NTU202403578

Figure 2.10 Schematic diagram of features in blood pressure

Furthermore, another previous study [10] indicated that Upstroke per Cardiac Cycle (UTCC) can also be a marker since the heart rate is highly differ from individuals to individuals.

$$UTCC = \frac{UT}{RRI} \tag{7}$$

where the RRI refers to the R-R Intervals from the measurement of ABP.

As the Figure 2.10 shown, the dicrotic notches are also our points of interests for involving feature selection. For this case, we adopted the quantity of both mean value of dicrotic notch amplitude and the time difference between systolic peak as our candidate of features. To avoid relying on both PPG and NBP records, which can introduce dependencies, we exclusively extract information about dicrotic notches from NBP signals. These notches are distinctly visible in schematic diagrams respectively, making them a reliable source for our analysis.



doi:10.6342/NTU202403578

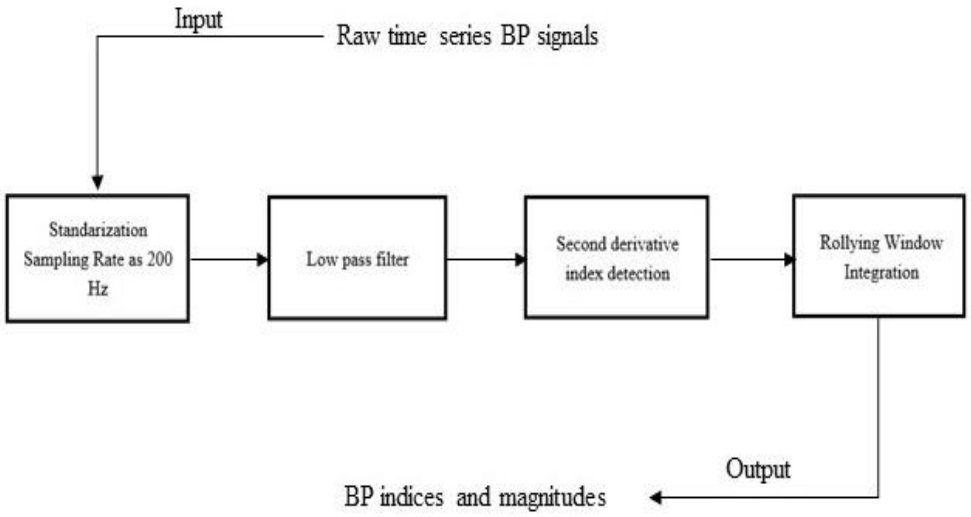


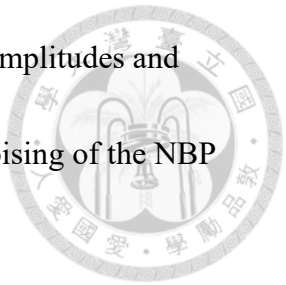

Figure 2.11 Flowchart of the Non-invasive Blood Pressure feature extraction

The feature extraction of NBP is a composition method as the flowchart Figure 2.11 displays in [35]. We employed a standardised procedure for processing NBP signals, considering their initial sampling frequency of 128 Hz, as provided by the IntelliVue Bedside patient monitor. The data were reshaped to a uniform frequency of 200 Hz to standardize the analysis. Subsequently, a zero-phase low-pass filter was applied to remove high-frequency noise components. The local minimums are the foot indices as well as the diastolic pressure. The second derivative of the signal was computed to identify inflection points i.e the Dicrotic notch points. Rolling Window Integration (RWI) was then applied to the second derivative to amplify peaks and improve the localisation accuracy of points of interest, which are the systolic pressure. This last



doi:10.6342/NTU202403578

step is meant to not only facilitated the feature extractions for both amplitudes and time duration but also contributed to the overall smoothing and denoising of the NBP signals.



## 2.6 Normalisation process

Since we have multiple features with very different magnitudes and distributions, the normalization is required before we adopt the Principal Component Analysis (PCA) we will apply later. One common way is to use the z-score normalization of each feature transforming them into scaled quantity. Scaling is a critical aspect of data analysis as it ensures that features are comparable. This process is essential for preventing features with larger scales from dominating the analysis and can enhance the performance and stability of machine learning algorithms. Furthermore, scaling aids in interpreting coefficients or feature importance, as it places all features on a similar scale.

$$z = \frac{x-\mu}{\sigma} \tag{6}$$

z: the new score after the features are transformed linearly

x: original value of each feature



doi:10.6342/NTU202403578

$\mu$: the mean of the original feature

$\sigma$: the SD of the original feature

After this scaling, newly scaled features are with characteristic of mean equals to zero and standard deviation equals to one.

# 2.7 Principal Component Analysis

The Table 2.5 is the selected features from both Time domain and Entropy domain; in total there are 25 and 3 features respectively. As we mentioned we exclude the Frequency domain features due to the similarity of them, they are not sufficiently distinct for effective classification in our study.



doi:10.6342/NTU202403578

| Domain | Features |
| --- | --- |
| Time Domain | Mean of RRI |
| | Mean of PTTp |
| | Mean of PTTs |
| | SDRR |
| | rmssd |
| | CoV. of PTTp |
| | CoV. of PTTs |
| | CoV. of RRI |
| | Mean of pNN50 |
| | age |
| | long-term mean of SBP |
| | long-term mean of DBP |
| | Mean of SBP |



doi:10.6342/NTU202403578

| | |
|---|---|
| Time Domain | Mean of AMP (Dicrotic amplitude) |
| | Mean of SUT |
| | mean of RRI (Systolic Intervals) |
| | SD of SBP |
| | SD of DBP |
| | SD of AMP (Dicrotic amplitude) |
| | SD of RRI (Systolic Intervals) |
| | SD of Systolic-Dicrotic Intervals |
| | SD of SUT |
| | mean of UTCC |
| | Mean of PP |
| | Mean of PP-Intervals |
| Entropy Domain | Shannon Entropy of RRI |
| | Complexity of SBP |



doi:10.6342/NTU202403578

| | |
|---|---|
| | Complexity of RRI (Systolic Intervals) |

Table 2.5 Selected features

Principal component analysis (PCA) is a common feature transformation method in Machine Learning. For the purpose of better detecting what features can explain the variance of the outcome and quantify the significance of every single original feature. It applies Single Value Decomposition (SVD) to the feature matrix. PCA reduces data dimensionality while keeping the majority of the original dataset's variety. It transforms the original features into a new set of orthogonal features (principal components), ordered by the amount of variance they explain.

PCA with SVD steps:

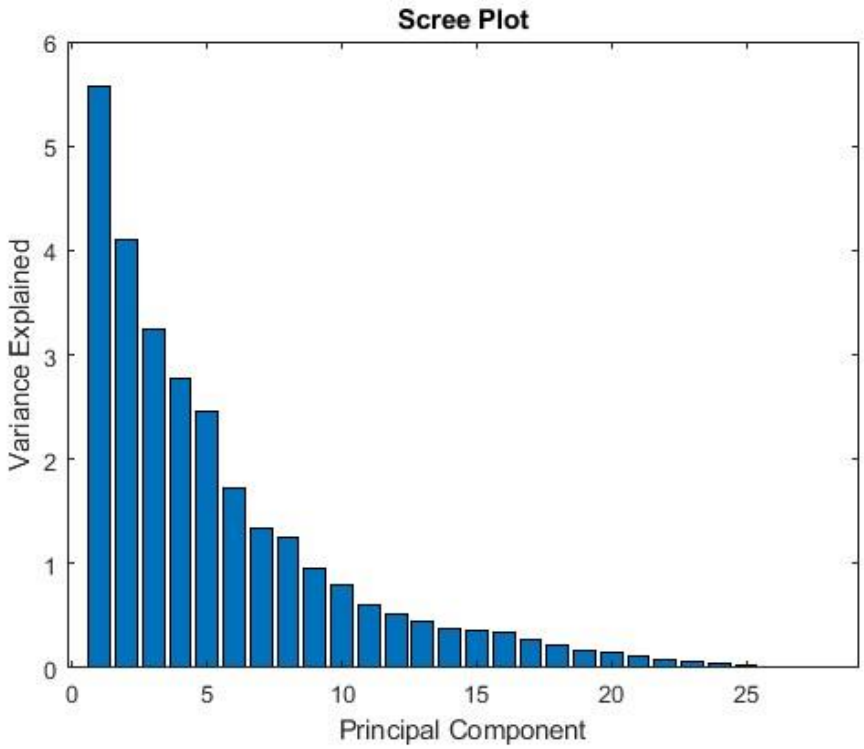
**Principal Component Analysis (PCA):**

- Centring the Data: Adjust data by subtracting mean values.
- Singular Value Decomposition (SVD): Decompose data matrix into orthogonal components U, D, and V.
- Principal Components (PCs): Directions maximising data variance, found in matrix V.



doi:10.6342/NTU202403578

- PC1 and PC2: First and second PCs capture highest orthogonal variance, forming new coordinate system.
- Additional PCs: Sequential components capture decreasing variance, each orthogonal to previous PCs.
- Key Concepts: Singular values quantify variance captured; eigenvalues indicate variance explained; loading scores define PC compositions.
- Scree Plot: Visualises variance explained by each PC, guiding component retention.






doi:10.6342/NTU202403578

Figure 2.12 Scree plot of PCs

Figure 2.12 presents the scree plot, depicting histograms that illustrate the distribution of principal components along with their corresponding eigenvalues. This plot aids in understanding how each principal component contributes to the overall variance explained in the dataset, helping to identify key components that significantly influence the data's structure.

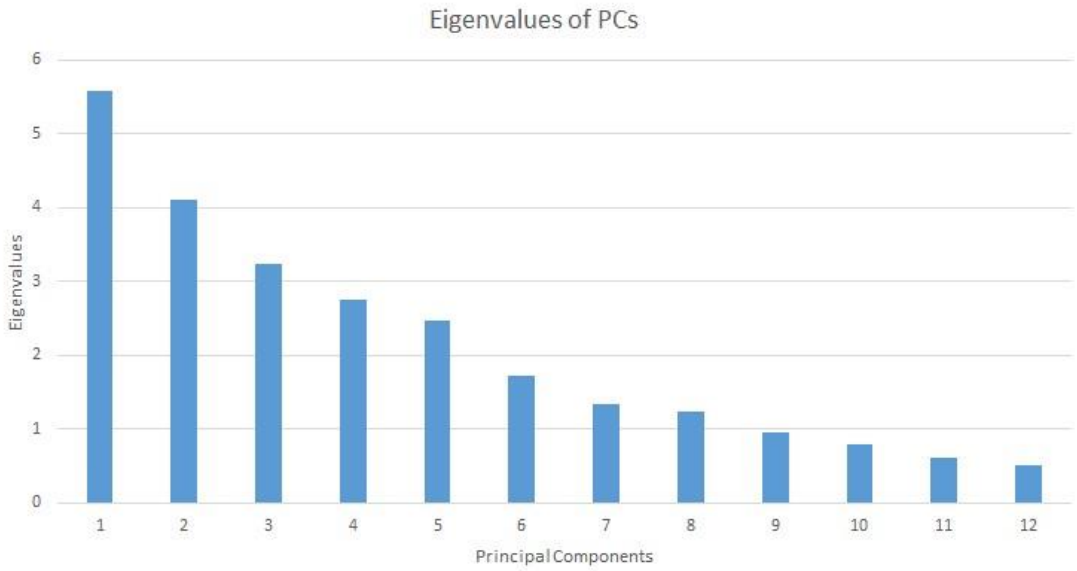



Figure 2.13 Eigenvalues of PCs

Table 2.5 lists every eigenvalue corresponding to the PC's. As the consequence, we can know there are eight eigenvalues of PC's are greater than 1.



doi:10.6342/NTU202403578

## 2.8 Feature Selection

To enhance methodological robustness and analytical clarity, a series of iterative refinements were undertaken in our study. In the entropy domain analysis, the selection of features necessitated careful consideration. Given the nuanced pathophysiological dynamics associated with stroke, particularly in elderly cohorts with pronounced cardiovascular comorbidities, the complexity of systolic pressure emerged as a salient candidate for feature extraction. This preference over diastolic pressure stemmed from the rationale articulated by [20], underscoring the pivotal role of systolic pressure in elucidating the intricate interplay between physiological parameters and pathological manifestations. Therefore, we have excluded the complexity of DBP from our feature list.

By emphasizing systolic pressure as a focal point of inquiry, the study endeavours to bridge the realms of clinical observation and physiological understanding. This strategic refinement holds promise for revealing subtle yet clinically significant variations in the studied population, thereby enriching the interpretative framework and advancing the overarching research objectives.

Coefficient of Variation (CoV.) is a statistical quantity to normalize the variations for



doi:10.6342/NTU202403578

each original feature in ML.

$$\text{CoV.} = \frac{SD}{mean} \quad (8)$$

SD: Standard Deviation of original features

mean: average values of original features

In our study, not all CoV. features will be included in the classification process, as some may pose challenges due to their irrelevant nature. Therefore, careful consideration and selection of relevant CoV. features are crucial for effective classification tasks.

# 2.9 Support Vector Machine Classification

Support Vector Machine (SVM) can be a practicable tool for classification in ML. It excels in handling data with high dimensionality and is computationally efficient, making it well-suited for supervised learning tasks. The reason why we choose it as the key component of algorithms is due to the nature small datasets alike the [1] implemented, It should be noted that there are many other algorithms that are applicable for smaller datasets, for example logistic regression.



doi:10.6342/NTU202403578

When it comes to our research, we first reiterate the basic math of SVM specifying the hyper-plane. The SVM algorithm tries to find the optimal hyper-plane that separates different classes of data points. The mathematical expression that defines this hyperplane is given by the following equation.

In the equation, $\vec{y}$ represents the output vector, which contains the predicted class labels. The term $\vec{w}$ is the weight matrix, which contains the coefficients that define the orientation of the hyperplane in the feature space. The input vector $\vec{x}$, consists of the data points we are trying to classify. Finally, $\vec{b}$ is the bias term, which adjusts the position of the hyperplane relative to the origin.

$$\vec{y} = \boldsymbol{w}^T\vec{x} + \vec{b} \tag{9}$$

The decision-making process in SVM involves calculating the dot product of the weight matrix w and the input vector $\vec{x}$, then adding the bias term $\vec{b}$. The data point is categorised into one class if $\vec{y}$ is greater than a particular threshold; if not, it is categorised into the other class.

This process ensures that the data points are separated by the maximum possible margin as Figure 2.14 shows the hyper-plain serve as the decision boundary, thereby improving the classification accuracy and robustness of the model.



doi:10.6342/NTU202403578

In supervised learning, the dataset is typically split into two subsets: the training set and the testing set. Here, we adopted a cross-validation technique to evaluate the model performance. The training set is used to train the AI model, where it learns the patterns and relationships in the data. The testing set is then used to evaluate the model's performance, assessing its generalization and accuracy. This process helps ensure that the model can make accurate predictions on new, unseen data. In brief, SVM aims to minimise the classification error, ensuring optimal performance in separating the data points.

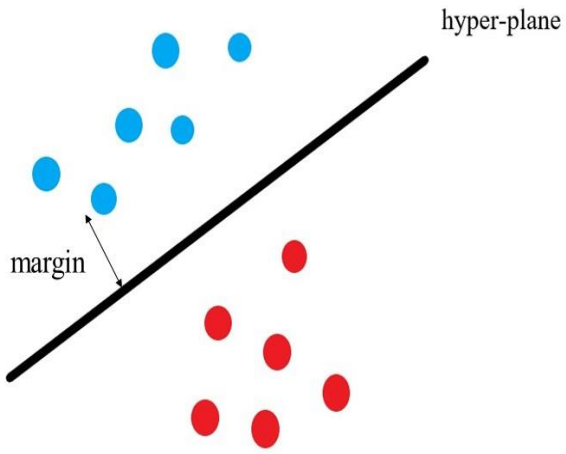



Figure. 2.14 schematic diagram of hyperplane in SVM



doi:10.6342/NTU202403578

A linear kernel function is a favoured and powerful option when it comes to transforming the raw data to subsequent analysis for its simplicity and convenience.

$$M=\begin{bmatrix} 0 & 4 & 10 \\ 4 & 0 & 10 \\ 10 & 4 & 0 \end{bmatrix}$$

The matrix M is our cost penalty matrix. We must made adjustments must to better fit a small population demographics and skewed target classes. A cost-penalty matrix is a technique used to adjust the classification process by assigning different costs or penalties to different types of classification errors. By incorporating a cost-penalty matrix, the classification algorithm can prioritize correct classification of certain classes over others, thus improving the overall performance of the classifier. The cost-penalty matrix is a square matrix used in machine learning, where the off-diagonal entries represent non-zero penalties assigned for misclassifications. In this context, it is logical that correctly classified instances, as represented by the diagonal entries, do not incur any penalty. Since the nature of imbalanced population for classification on my topic, the cost-penalty matrix is incorporated in our classification task to perform a more desirable outcome for certain topics.



doi:10.6342/NTU202403578

# 2.10 Conclusions

In this chapter, we reviewed some mathematical quantity and certain morphology of the physiological measurements as features aiming to distinguish the different extents of impairment when measurable cardiovascular indicators are quantified properly which may provide sufficient information for prognosis. We will reveal the evaluation of model performance and generalisation ability later on.



doi:10.6342/NTU202403578

# Chapter 3

# Results and Discussions

In this chapter, we display the classification results and analysis to have a deeper insight on how ML make decision by our designated process. As we exclude some less relevant features of our classification in the previous chapter, we can obtain more desirable outcome after ML implementation.

## 3.1 Classification Discussions

### 3.1.1. Data Splitting

The dataset is divided into training and testing sets, with a 65% portion allocated for training and a 35% portion for testing randomly. This ensures that the model is trained on a subset of the data and evaluated on unseen data to assess its generalisation performance and reproducibility

### 3.1.2. Training and Testing

The training set is used to train the machine learning model, which involves learning



doi:10.6342/NTU202403578

the underlying patterns and relationships in the data. In this script, the training set consists of various features extracted from clinical data, such as physiological measurements and patient demographics PCA is applied to reduce the dimensionality of the input features and employ the most significant components to trade-off between adaptability of implementation and The SVM model is then trained on the transformed features to classify patients into different severities based on their modified mRS scores. After this step, we keep testing set is separated from the training set and is meant to evaluate the performance of the trained model.

### 3.1.3. Evaluation

Performance metrics are computed to assess the effectiveness of the trained model in predicting patient outcomes. The confusion matrix, which provides the comparison of predicted and actual class labels.

We compressed the mRS scale with 0-2, 3-4 and 5-6 respectively since the diagnosis cannot be precisely given by medical doctors due to some transition state at that moment when patients were sent to ICU. For example, when diagnosing certain patients with mRS scores between 3 and 4, we categorise these cases as mRS 3 for a



doi:10.6342/NTU202403578

slightly less severe one.

Initially, we included the entire population from the ICU with 221 patients in total. However, some of their real-time NBP data was not properly recorded. For this reason, we only adopted ECG and PPG, along with long-term average blood pressure for both systolic and diastolic pressure from tables provided by the hospital. The preliminary results were poor, with an accuracy of 52.727% and an AuROC of 0.68.

After discussing with a clinical physician, he suggested focusing on the patients with complete NBP records. The waveform of dynamic blood pressure can aid our study, as it reflects the circulation of blood from the heart to the brain. Consequently, we investigated the subsets with complete records. For severe cases, ECG can only exhibit arrhythmia and provides little information on the damaged cerebral situation.

For smaller datasets, feature engineering is critical for ML models. We refined the process of feature extraction by analysing the patterns of PPG in addition to Pulse Transit Time. After conducting all standardised procedures of the algorithms sequentially, we obtained the optimised classification results, displayed in Figure 3.1.



doi:10.6342/NTU202403578

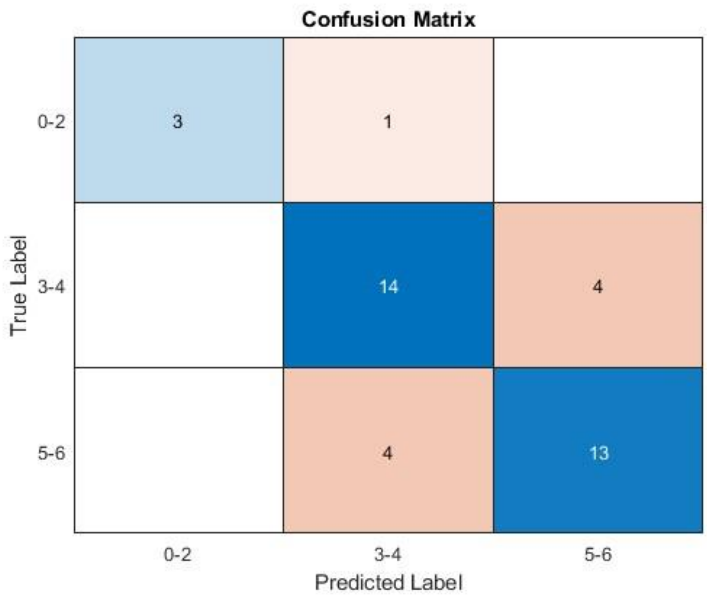



Figure. 3.1. Confusion matrix of the testing patients

The testing patients are in total 39 of them which were split randomly from original populations. The diagonal entry of the confusion matrix is the correctly classified and the others are incorrectly classified. As Figure 3.1, we originally have a relatively skewed distribution from our patients of interests.

For our classification task, we retained the first to twelfth principal components to keep the majority information of PCs as Figure 3.2. Subsequently, we employed a cost-sensitive SVM approach to visualise the performance.



doi:10.6342/NTU202403578

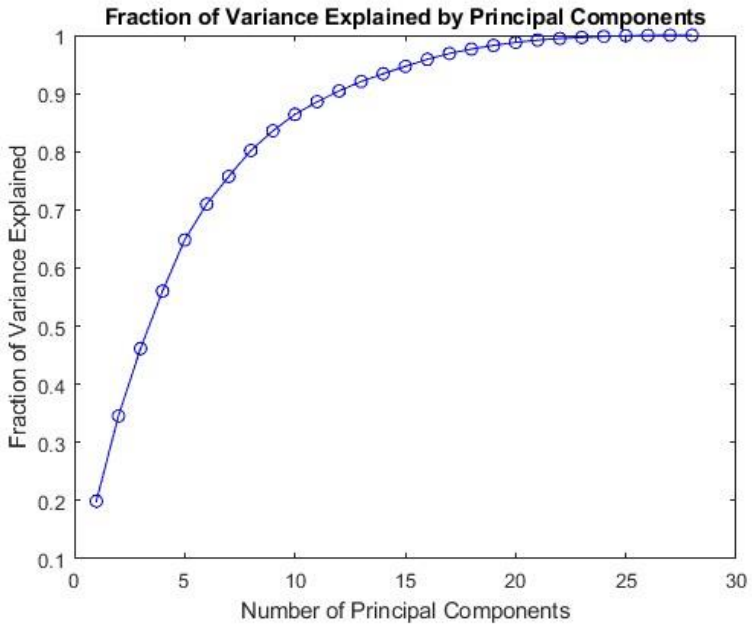



Figure 3.2 Fraction of Variance plots with increased Principal Components

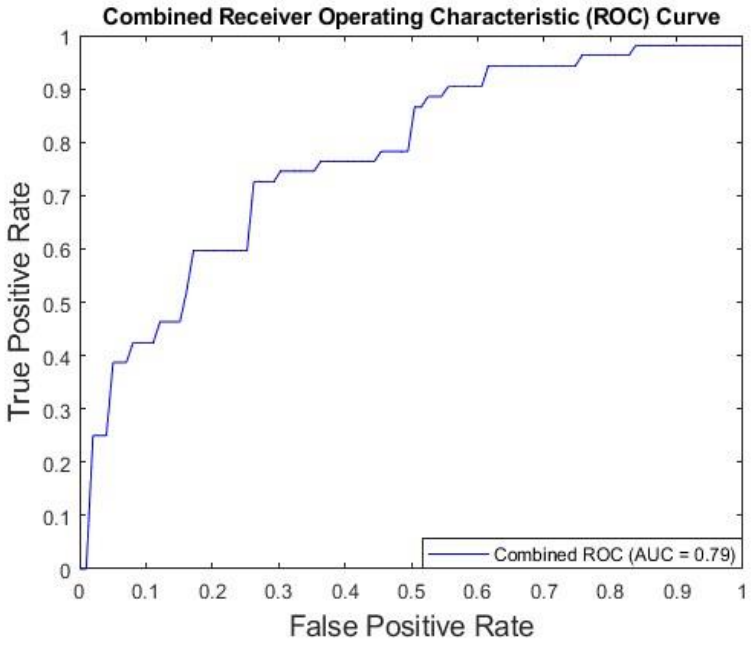



Figure 3.3 ROC Curve of classification



doi:10.6342/NTU202403578

| | |
|---|---|
| Mean of SBP long-term | 3.6841 |
| Shannon_En of RRI | 3.6840 |
| UTCC | 3.6327 |
| Complexity of RRI | 3.5931 |
| Mean of DBP long-term | 3.5175 |
| SD of Systolic-Dicrotic Interval | 3.5158 |
| SD of SBP | 3.4906 |
| CoV of PTTs | 3.4706 |
| Mean of Dicrotic Amplitude | 3.4648 |
| pNN50 | 3.4380 |
| SD of RRI(Systolic Interval) | 3.4031 |
| Mean of PTTs: | 3.3902 |
| SD of DBP: | 3.3883 |
| Mean of RRI | 3.3677 |
| CoV of PTTp | 3.3587 |
| SD of Dicrotic Amplitude | 3.3114 |
| age | 3.3079 |
| SD of SUT | 3.2573 |
| Mean of PP | 3.2466 |
| Mean of RRI(Systolic Interval) | 3.2106 |
| SD of SUT: | 3.2102 |
| Mean of PP-Interval | 3.0607 |
| Mean of PTTp | 2.9342 |
| Complexity of SBP: | 2.9141 |
| CoV of RRI | 2.8359 |
| rmssd | 2.7870 |
| SDRR | 2.6933 |
| Mean of SBP | 2.6424 |

Table 3.1 absolute value loading summation of original features over every single PC

The final results demonstrate an accuracy of 0.76923 and an AuROC of 0.79 as Figure 3.3 shown. After computation, we gained the table 3.1 indicates that features ranked



doi:10.6342/NTU202403578

by the absolute values of the loading summation of each Principal Component (PC) reveal that the long-term average systolic blood pressure, Shannon Entropy of RR-Intervals, and UTCC have the highest significance for the classification model. Additionally, the complexity of systolic pressure also contributes significantly. However, the small differences between these values present an initial challenge. As we mentioned, features from the Frequency domain were excluded due to their relatively similar values across the population.

## 3.2 Conclusions

In this chapter, we follow the previous chapter in terms of physiological features and algorithms aiming for a desirable ML model. Thus, we present the confusion matrix and ROC curve as our primary focus for evaluating the model on testing patients split from our filtered dataset. Given the complexity of unknown pathologies, the role of physiological features and the careful tailoring of input are crucial in bridging the gap between ML models and potential pathology, as acknowledged by medical scholars and professionals.



doi:10.6342/NTU202403578

Incorporating hospital-provided features such as age and long-term blood pressure into our study has provided valuable insights into physiological patterns. However, the challenge remains in utilising binary data for factors like smoking and diabetes, where quantifying severity without detailed pathology information proves difficult. Addressing these complexities requires further exploration of suitable methodologies to enhance the robustness of our findings.



doi:10.6342/NTU202403578

# Chapter 4

# Conclusions and Future Work

## 4.1 Summary of Findings

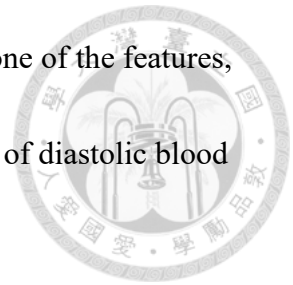
Our study initially focused on ECG and HRV analysis, but we faced challenges with missed blood pressure data among our original population. After securing real-time blood pressure signals, we've found that analysing Systolic Upstroke Time (SUT) and the dicrotic notch for both amplitude and time intervals following each systolic peak enhances the feasibility and robustness of our classification. Long-term mean pressure measurements for both systolic and diastolic values provide valuable insights into physiological conditions.

Some patients on the verge of shock may exhibit very slow heart rates and unusual hypotension. In our study, other features from PPG play a crucial role in ML. Despite the limitations of PCA, which doesn't explicitly highlight critical features, we advocate for including all three measurements to enhance both accuracy and the generalisation capability of our classification efforts.

For performance evaluation, it aligns with mainstream medical opinion that systolic pressure is more significant than diastolic pressure, particularly in Entropy Domain



doi:10.6342/NTU202403578



analysis. By selecting the complexity of systolic blood pressure as one of the features, the classification is more separable than when using the complexity of diastolic blood pressure or putting them all together as features.

## 4.2 Limitations

Our research aims to deepen the understanding of the correlation between stroke severity and cardiovascular health. Signal morphology analysis, particularly within the time domain, is crucial due to its simplicity and intuitiveness. However, more advanced methodologies are needed to address confounding factors and improve robustness.

To mitigate artifacts, we propose moving beyond standardised filtering to adaptive strategies that adjust parameters dynamically. ML techniques can also help distinguish true physiological signals from artifacts, enhancing analysis fidelity. Human factors can significantly influence our research direction. Personalised analyses considering patient demographics and clinical histories can uncover nuanced insights into signal morphology and cardiovascular health. Collaboration with clinical experts adds valuable domain knowledge to our investigations.



doi:10.6342/NTU202403578

Moreover, external validation with independent datasets, are essential to ensure reliability and generalisability. Longitudinal studies tracking changes in signal patterns and health outcomes over time offer valuable insights into the progression of stroke severity and cardiovascular health.

By refining our methodologies and acknowledging the multifaceted nature of our research, we work on significantly contributing to the understanding of stroke severity and its relationship with cardiovascular health.

Given that we have just over a hundred valid patient samples from the Intensive Care Unit (ICU) of the Department of Neurology at NTUH, our Machine Learning model is applicable primarily to ischaemic embolism stroke situations. For other similar pathological conditions, we must rely on additional studies to support certain assertions, due to the inherent constraints of this study. Medical professionals may contribute more primitive studies rather than a statistical AI modeling by rigid experiments on humans and Anatomy is the key to investigate more sophisticated nature of life science studies.

Additionally, several lifestyle factors, such as smoking, alcohol consumption, regular exercise, adequate sleep and a healthy diet, can influence the risk of AF and other



doi:10.6342/NTU202403578

cardiovascular diseases. These lifestyle details are typically reported by patients themselves rather than by medical professionals. Patients and care givers should be always aware of their lifestyle habits to prevent deteriorating situations.

## 4.3 Future Work

As noted in [1], a successful model can distinguish between mRS 0-2 and 3-6 in patients with the same described pathology conditions. Building on this model, we refined certain aspects of filtering and algorithms to further separate mRS 3-4 and 5-6 using a ML model. We could improve this model by reiterating the feature selection and denoising by other method for instance detrending analysis due to the baseline wandering effect which can drastically influence area integration analysis.

If we are being able to include more patients from different races and countries by global collaboration, we could make improvement on both algorithm selection and learning dataset. However, to gain deeper insights into cardiology and vascular disease-related physiological indices, and to observe more intriguing phenomena more detailed collaboration with medical doctors and other professionals, such as biomedical engineers, is essential.



doi:10.6342/NTU202403578

This collaboration is necessary to develop rapid and portable equipment like wearable devices for instant monitoring of at-risk patients. Such equipment can provide early warnings to prevent severe incidents, reducing the need for intense medical treatment and rehabilitation. We are still on the way to promote and popularise the effective and inexpensive monitoring outside of medical institutions.



doi:10.6342/NTU202403578

# References

[1] Ding-Yuan Lee et al, Predicting Stroke Outcomes in Atrial Fibrillation Patients Using Multimodal Analysis of Physiological Signals, 2015

[2] Sung-Chun Tang et al, Complexity of heart rate variability predicts outcome in intensive care unit admitted patients with acute stroke, 2014

[3] Chih-Hao Chen et al, Complexity of Heart Rate Variability Can Predict Stroke-In-Evolution in Acute Ischemic Stroke Patients, 2015

[4] Sung-Chun Tang et al, Identification of Atrial Fibrillation by Quantitative Analyses of Fingertip Photoplethysmography, 2017

[6] Jiapu Pan and Willis J. Tompkins, A Real-Time QRS Detection Algorithm 1985

[7] Yun-Kai Lee et al, Blood Pressure Complexity Discriminates Pathological Beat-to-Beat Variability as a Marker of Vascular Aging, 2022

[8] Tatsuya Maruhashi et al,Upstroke Time Is a Useful Vascular Marker for Detecting Patients With Coronary Artery Disease Among Subjects With Normal Ankle-Brachial Index, 2020 :1-2

[9] Soler EP, Ruiz VC. Epidemiology and risk factors of cerebral ischemia and



doi:10.6342/NTU202403578

ischemic heart diseases: similarities and differences. Curr Cardiol Rev. 2010 Aug;6(3):138-49. doi: 10.2174/157340310791658785. PMID: 21804773; PMCID: PMC2994106. 2010

[10] Po-Chao Hsu et al, Upstroke Time Per Cardiac Cycle as A Novel Parameter for Mortality Prediction in Patients with Acute Myocardial Infarction, 2020

[11] Nayak S, Natarajan B, Pai RG. Etiology, Pathology, and Classification of Atrial Fibrillation. 2020

[12] Fred Shaffer and J. P. Ginsberg, An Overview of Heart Rate Variability Metrics and Norms 2017

[13] Roy M. John(PhD)Saurabh Kumar(PhD), Sinus Node and Atrial Arrhythmias. 2016

[14] Marit H. N. van Velzen, Increasing accuracy of pulse transit time measurements by automated elimination of distorted photoplethysmography wave 2017

[15]https://www.cdc.gov/heartdisease/atrial_fibrillation.htm

[16]https://www.melbourneheartrhythm.com.au/learn/conditions/73-atrial-fibrillation

[17]https://www.richtek.com/Design%20Support/Technical%20Document/AN057?sc



doi:10.6342/NTU202403578

_lang=zh-TW

[18] Tanja Charlotte Frederiksen, The bidirectional association between atrial fibrillation and myocardial infarction 2023

[19]https://www.physio-pedia.com/Middle_Cerebral_Artery

[20] Basile JN. Systolic blood pressure. BMJ. 2002 Oct 26;325(7370):917-8. doi: 10.1136/bmj.325.7370.917. PMID: 12399325; PMCID: PMC1124431.

[21] Yihong Sun. The link between diabetes and atrial fibrillation: cause or correlation? 2009

[22] João D. Fontes .Insulin Resistance and Atrial Fibrillation (from the Framingham Heart Study) 2012

[23] Nogles TE, Galuska MA. Middle Cerebral Artery Stroke. 2023 Apr 3. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan–. PMID: 32310592.

[24]https://www.epilepsysparks.com/brain-lobes

[25]https://www.firstaidforfree.com/recording-a-12-lead-ecgekg/

[26]https://www.biometriccables.in/blogs/blog/12-lead-ecg-cable-electrode-placement



doi:10.6342/NTU202403578

[27] *Oh, G.C., Cho, HJ*. Blood pressure and heart failure. Clin Hypertens 26, 1 (2020). https://doi.org/10.1186/s40885-019-0132-x

[28] Karin Willeit, Stefan Kiechl, Metrics Atherosclerosis and atrial fibrillation – Two closely intertwined diseases 2013

DOI:https://doi.org/10.1016/j.atherosclerosis.2013.11.082

[29] Li C, Wang H, Li M, Qiu X, Wang Q, Sun J, Yang M, Feng X, Meng S, Zhang P, Liu B, Li W, Chen M, Zhao Y, Zhang R, Mo B, Zhu Y, Zhou B, Chen M, Liu X, Zhao Y, Shen M, Huang J, Luo L, Wu H, Li YG. Epidemiology of Atrial Fibrillation and Related Myocardial Ischemia or Arrhythmia Events in Chinese Community Population in 2019. Front Cardiovasc Med. 2022 Apr 4;9:821960. doi: 10.3389/fcvm.2022.821960. PMID: 35445083; PMCID: PMC9013769.

[30] Ding X, Zhang YT. Pulse transit time technique for cuffless unobtrusive blood pressure measurement: from theory to algorithm. Biomed Eng Lett. 2019 Feb 18;9(1):37-52. doi: 10.1007/s13534-019-00096-x. PMID: 30956879; PMCID: PMC6431352.

[31] Chen H, Chen G, Zhang L, Wu W, Li W, Wang X, Yan X, Chen Y, Wu S. Estimated pulse wave velocity can predict the incidence of new-onset atrial



doi:10.6342/NTU202403578

fibrillation: A 11-year prospective study in a Chinese population. Front Cardiovasc Med. 2022 Aug 22;9:912573. doi: 10.3389/fcvm.2022.912573. PMID: 36072866; PMCID: PMC9443485.

[32] Mitchell GF, Vasan RS, Keyes MJ, et al. Pulse Pressure and Risk of New-Onset Atrial Fibrillation. *JAMA.* 2007;297(7):709–715. doi:10.1001/jama.297.7.709

[33]https://en.wikipedia.org/wiki/Lobes_of_the_brain

[34]https://mdmedicine.wordpress.com/2011/04/24/heart-conduction-system/

[35]https://www.mathworks.com/matlabcentral/fileexchange/60172-bp_annotate.

[36] Cepelis A, Brumpton BM, Malmo V, et al. Associations of Asthma and Asthma Control With Atrial Fibrillation Risk: Results From the Nord-Trøndelag Health Study (HUNT). *JAMA Cardiol.* 2018;3(8):721–728. doi:10.1001/jamacardio.2018.1901

[37] Hwang, C.S.; Kim, Y.H.; Hyun, J.K.; Kim, J.H.; Lee, S.R.; Kim, C.M.; Nam, J.W.; Kim, E.Y.,Evaluation of the Photoplethysmogram-Based Deep Learning Model for Continuous Respiratory Rate Estimation in Surgical Intensive Care Unit. *Bioengineering* 2023, *10*, 1222. https://doi.org/10.3390/bioengineering10101222



doi:10.6342/NTU202403578

[38]https://thoracickey.com/the-electrocardiogram-2/

[39] Block, R.C. Yavarimanesh, M., Natarajan, K. et al. Conventional pulse transit times as markers of blood pressure changes in humans. *Sci Rep* **10**, 16373 (2020). https://doi.org/10.1038/s41598-020-73143-8

[40] Hrabia JB, Pogue EPL, Zayachkowski AG, Długosz D, Kruszelnicka O, Surdacki A, Chyrchel B. Left atrial compliance: an overlooked predictor of clinical outcome in patients with mitral stenosis or atrial fibrillation undergoing invasive management. Postepy Kardiol Interwencyjnej. 2018;14(2):120-127. doi: 10.5114/aic.2018.76402. Epub 2018 Jun 19. PMID: 30008763; PMCID: PMC6041835.

[41] García-Escobar A, Vera-Vera S, Jurado-Román A, Jiménez-Valero S, Galeote G, Moreno *R*. Subtle QRS changes are associated with reduced ejection fraction, diastolic dysfunction, and heart failure development and therapy responsiveness: Applications for artificial intelligence to ECG. Ann Noninvasive Electrocardiol. 2022 Nov;27(6):e12998. doi: 10.1111/anec.12998. Epub 2022 Jul 29. PMID: 35904538; PMCID: PMC9674781.

[42] Zuo K, Li K, Liu M, Li J, Liu X, Liu X, Zhong J, Yang X. Correlation of left atrial wall thickness and atrial remodeling in atrial fibrillation: Study based on low-



doi:10.6342/NTU202403578

dose-ibutilide-facilitated catheter ablation. Medicine (Baltimore). 2019 Apr;98(15):e15170. doi: 10.1097/MD.0000000000015170. PMID: 30985700; PMCID: PMC6485781.

[43] Hiraga A, Yamaoka T, Sakai Y, Osakabe Y, Suzuki A, Hirose N. Relationship between outcome in acute stroke patients and multiple stroke related scores obtained after onset of stroke. J Phys Ther Sci. 2018 Oct;30(10):1310-1314. doi: 10.1589/jpts.30.1310. Epub 2018 Oct 12. PMID: 30349170; PMCID: PMC6181659.

[44]https://www.strokeinfo.org/stroke-facts-statistics/

[45] Andreozzi E, Sabbadini R, Centracchio J, Bifulco P, Irace A, Breglio G, Riccio M. Multimodal Finger Pulse Wave Sensing: Comparison of Forcecardiography and Photoplethysmography Sensors. *Sensors*. 2022; 22(19):7566. https://doi.org/10.3390/s22197566

[46]https://www.thelancet.com/journals/lanepe/article/PIIS2666-7762(23)00205-3/fulltext

[47]https://manual.jointcommission.org/releases/TJC2018A/DataElem0569.html

[48]https://www.mesa-nhlbi.org/ParticipantWebsite/MesaNewsAirPollutionsCarotid.aspx



doi:10.6342/NTU202403578

[49]https://www.radiologymasterclass.co.uk/tutorials/ct/ct_acute_brain/ct_brain_acute_ischaemia

[50]https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html

**Acknowledgements:**

This work was supported by: National Taiwan University Hospital, research grants NTUH PC851



doi:10.6342/NTU202403578