Automatic Cardioembolic Stroke Prediction using Clinical Features and Non-contrast CT Images

Mr. Pasit Jakkrawankul

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2021

ระบบอัตโนมัติสำหรับการประเมินความเสี่ยงโรคหลอดเลือดสมองอุดตันจากลิ่มเลือดหัวใจโดยใช้ข้อมูลทางคลินิกและภาพถ่ายซีทีแสกนสมองปกติ

นายพสิษฐ์ จักรวาลกุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Thesis Title	Automatic Cardioembolic Stroke Prediction using Clinical Features and Non-contrast CT Images
By	Mr. Pasit Jakkrawankul
Field of Study	Computer Science
Thesis Advisor	Doctor Ekapol Chuangsuwanich, Ph.D.
Thesis Co Advisor	Associate Professor Proadpran Punyabukkana, Ph.D.

Accepted by the FACULTY OF ENGINEERING, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

..... Dean of the FACULTY OF ENGINEERING

()

THESIS COMMITTEE

..... Chairman

(Associate Professor THANARAT CHALIDABHONGSE, Ph.D.)

..... Thesis Advisor

(Doctor Ekapol Chuangsuwanich, Ph.D.)

..... Thesis Co-Advisor

(Associate Professor Proadpran Punyabukkana, Ph.D.)

..... Examiner

(Doctor Chaipat Chunharas, M.D.)

..... External Examiner

(Assistant Professor Theerawit Wilaiprasitporn, Ph.D.)

พสิษฐ์ จักรวาลกุล : ระบบอัตโนมัติสำหรับการประเมินความเสี่ยงโรคหลอดเลือดสมองอุดตันจากลิ่มเลือดหัวใจโดยใช้ข้อมูลทางคลินิกและภาพถ่ายซีทีแสกนสมองปกติ. (Automatic Cardioembolic Stroke Prediction using Clinical Features and Non-contrast CT Images) อ.ที่ปรึกษาหลัก : อ. ดร.เอกพล ช่วงสุวนิช, อ.ที่ปรึกษาร่วม : รศ. ดร.โปรดปราน บุณยพุกกณะ

โรคหลอดเลือดสมองอุดตันจากลิ่มเลือดหัวใจเป็นโรคหลอดเลือดสมองตีบประเภทหนึ่งที่มีความอันตรายอย่างมาก ผู้ป่วยที่เป็นโรคหลอดเลือดสมองชนิดนี้ต้องการการรักษาที่เฉพาะเจาะจงเพื่อป้องกันไม่ให้เกิดการอุดตันขึ้นอีก การป้องกันนั้นมีความสำคัญอย่างยิ่งเนื่องจากการอุดตันของโรคหลอดเลือดสมองชนิดนี้ก่อให้เกิดความเสียหายต่อเนื้อสมองเป็นบริเวณกว้าง ดังนั้นการจำแนกประเภทของโรคหลอดเลือดสมองตีบชนิดนี้ออกจากประเภทอื่นๆ จึงเป็นสิ่งสำคัญ เราจึงพัฒนาโมเดลปัญญาประดิษฐ์ที่สามารถวิเคราะห์ทั้งข้อมูลทางคลินิกขั้นพื้นฐานและภาพถ่าย CT แบบปกติเพื่อทำนายความเสี่ยงของโรคหลอดเลือดสมองอุดตันจากลิ่มเลือดหัวใจ ประสิทธิภาพของวิธีการของเราซึ่งวัดด้วยพื้นที่ภายใต้กราฟ receiver operating characteristic curve (ROC-AUC) นั้นอยู่ที่ 0.840 โดยใช้ชุดข้อมูลของผู้ป่วยโรคหลอดเลือดสมองเพียง 227 ตัวอย่าง นอกจากความสามารถในการจำแนกประเภทย่อยของโรคหลอดเลือดสมองตีบแล้ว เรายังสามารถระบุบริเวณที่สมองขาดเลือดและความสำคัญของอาการทางคลินิกได้อีกด้วย นอกจากนั้น วิธีการของเราสามารถนำมาใช้ได้อย่างกว้างขวาง เนื่องจากเราต้องการเพียงข้อมูลทางคลินิกขั้นพื้นฐานและการตรวจ CT แบบปกติซึ่งมีอยู่ในโรงพยาบาลทั่วไป

สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อนิสิต
		.....
ปี	2564	ลายมือชื่อ อ.ที่ปรึกษาหลัก
การศึกษา		.....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม

6272058721 : MAJOR COMPUTER SCIENCE

KEYWORD:

Pasit Jakkrawankul : Automatic Cardioembolic Stroke Prediction using Clinical Features and Non-contrast CT Images. Advisor: Dr. Ekapol Chuangsuwanich, Ph.D. Co-advisor: Assoc. Prof. Proadpran Punyabukkana, Ph.D.

Cardioembolic stroke is a dangerous subtype of ischemic stroke. The patients with this subtype need special treatments to prevent recurrent events. The prevention is vital since only one more event could result in fatal damage. Hence, the classification into the categories of cardioembolic and non-cardioembolic subtypes is essential. We developed a multimodal machine learning model that can integrate the basic clinical information and non-contrast CT images to predict the risk of cardioembolic stroke. Our method reached the areas under the receiver operating characteristic curve (ROC-AUC) of 0.840 by using a dataset of only 227 samples of stroke patients. Besides the capability to classify the stroke subtypes, the method can provide the interpretability of the model decision in the forms of the heatmap for large infarct localization and the feature impacts for interpretation. Our approach can be widely applied since we need only the basic clinical information and non-contrast CT which are commonly available in general hospitals.

Field of Study:	Computer Science	Student's Signature .....
Academic Year:	2021	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

This work is supported by Chulalongkorn University Technology Center. I would like to thank to Chaipat Chunharas, Wasan Akarathanawat, Pongpat Vorasayan, and Sedthapong Chunamchai from King Chulalongkorn Memorial Hospital for their generous effort in data collection and data labelling. Moreover, I would like to thank to Ekapol Chuangsuwanich, Proadpran Punyabukkana, and Naruemon Pratanwanich from Chulalongkorn University for suggestions and technical knowledges.

Pasit Jakkrawankul

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. Introduction

1.1.Motivation

Stroke is one of the most severe diseases causing serious disability or even death worldwide. To provide effective treatment to stroke patients, it is essential to identify the etiological subtype. Cardioembolic is a dangerous subtype leading to extensive damage to the brain tissue. Consequently, the main focus of the treatment for this subtype is to prevent recurrent events, since the secondary damage could be fatal. Anticoagulant is an effective therapy for the cardioembolic subtype, it can prevent the development of the malicious embolus. However, it has negative effects on the other subtypes, thus distinguishing cardioembolic from the others is critical for stroke treatment. Traditionally, classifying the subtype of stroke is time-consuming, especially monitoring the electrocardiogram (EKG) for patients with suspected cardioembolism. Therefore, clinicians are responsible to prioritize the risk of cardioembolic stroke to optimize the investigation resources. Nevertheless, the risk estimation is challenging for general practitioners. In many cases, they need advice from experts for consideration. This could be a disadvantage for the hospitals with limited resources such as rural hospitals with a shortage of stroke specialists. Fortunately, with the rise of modern machine learning technologies, various automated assisting systems have been developed to overcome the challenges.

In stroke diagnosis, there are two basic sources of information. The first is clinical information including age, vital signs, symptoms, underlying disease, and physical examination results. The other source is non-contrast computed tomography (CT) which is useful for detecting diseases in the brain. With the advances in machine learning techniques, these sources of information can be mutually analyzed to classify the subtypes of stroke. Besides the accurate risk estimation results, the interpretability of the machine learning model's decision also plays a key role to enable trustworthy assistance in clinical diagnosis.

In this study, we aim to develop a multimodal method that combines clinical information and CT to predict the risk of cardioembolic stroke. Non-contrast CT images are used to determine the existence of large infarct which is a characteristic of cardioembolic stroke. We apply a novel deep learning technique, pyramid localization network (PYLON) architecture (Preechakul, Sriswasdi et al. 2020) that specializes in weakly supervised localization for medical images, to simultaneously predict the likelihood and identify the location of large infarcts in CT images. The predicted likelihoods were utilized as the image features together with the clinical features extracted from electronic health records (EHR). These features are then processed by machine learning methods to estimate the risk of cardioembolic stroke.

1.2. Objective

We aim to develop an automated system to estimate the risk of cardioembolic stroke using both clinical information and CT images. The main hypothesis of this study is:

The characteristics of cardioembolic stroke can be explained by the clinical information and the brain CT findings. Therefore, machine learning techniques can be applied to analyze the information and estimate the risk of cardioembolic stroke.

The objectives of this thesis are as follows.

1. To develop a multimodal machine learning method to integrate the clinical information and CT images to predict the risk of cardioembolic stroke.
2. To enable the interpretability of the model's decision for better understanding and reliable prediction results.

1.3. Scope

The scope of this thesis is to apply machine learning techniques with the clinical information and CT images extracted from King Chulalongkorn Memorial Hospital (KCMH) to estimate the risk of cardioembolic stroke together with the model interpretability.

2. Related work

Ischemic stroke subtype classification had been studied in various aspects (Amarenco, Bogousslavsky et al. 2009), (Amort, Fluri et al. 2012). Oxfordshire Community Stroke Project (OCSP) (Bamford, Sandercock et al. 1991) categorizes ischemic stroke into 4 phenotypes which include total anteriorcirculation infarcts (TACI), lacunar infarcts (LACI), partial anterior circulation infarcts (PACI), and posterior circulation infarcts (POCI). Despite the simplicity of the system, only infarct location is not sufficient to specify a cardiac source of embolism. Trial of Org10172 in Acute Stroke (TOAST) classification (HP, BH et al. 1993) is the most wildly used method due to its simplicity and the ability to specify stroke etiology. It classifies stroke subtypes into 5 categories including large-artery atherosclerosis (LAA), cardioembolism, small-vessel occlusion, a stroke of other determined etiology, and stroke of undetermined etiology. We applied this system to categorize stroke patients into the 5 groups and used them as the ground truths for our study. Causative Classification System (CCS) (Ay, Benner et al. 2007) is another system that divides ischemic stroke into 5 subtypes similar to TOAST. The undetermined group is further divided into cryptogenic embolism, other cryptogenic, incomplete evaluation, and unclassified categories. Moreover, each group is subdivided based on the reliability of evidence as evident, probable or possible. Atherosclerosis, Small-vessel disease, Cardioembolism and Other cause (ASCO) system (Amarenco, Bogousslavsky et al. 2013) categorizes subtypes by causes as follows: atherosclerosis, small vessel disease, cardiac source, and other cause. Also, the presence and causal relationship to the stroke is graded for each potential cause. However, these two systems are more complicated compared to TOAST and not widely adopted.

Recently, these stroke subtyping processes have been transformed into automated systems with the advance in machine learning techniques. Fang et al. (Fang, Xu et al. 2020) developed an automated ischemic stroke subtyping using OCSP as the gold standard. They studied the effectiveness of numerous machine learning models on a dataset from The International Stroke Trial (IST). The dataset of 16,636 patients was used

to select robust features by applying recursive feature elimination (RFE) incorporated with linear Support Vector Classifier (SVC), Random Forest, Extra Tree Classifier, Adaboost, and Multinomial Naïve Bayes. Then, another Extra Tree Classifier and a simple neural network were used to classify the stroke phenotypes. They achieved over 0.95 accuracy with both classifiers.

Clinical feature extraction from EHR is a key challenge for the development of machine learning because of the unstructured nature of free text. Fortunately, the recent advance in natural language processing (NLP) had successfully eliminated the barrier and made the task a lot more practical. Garg et al. (Garg, Oh et al. 2019) proposed an automating ischemic stroke subtype classification using NLP to extract clinical features from EHR of 1,091 patients. TOAST classification was used as the ground truth for this study. They experimented with 7 machine learning techniques including K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Extra Tree Classifier, Gradient Boosting Tree, XGBoost, and stacking of these models using logistic regression. The stacking model attained the best performance with kappa coefficient of 0.72. Sung et al. (Sung, Lin et al. 2020) employed NLP to extract features from EHR of 4,640 patients and then combine the features with National Institutes of Health Stroke Scale (NIHSS) to form the initial features. A correlation-based feature selection technique was applied to reduce the feature's dimension. Six commonly used machine learning techniques were utilized to classify the OCSP subtypes. The best method reached 0.399 Cohen's kappa coefficient. Guan et al. (Guan, Ko et al. 2020) proposed an automated electronic phenotyping of cardioembolic stroke. They developed an NLP algorithm to extract cardioembolic stroke features using text-mining on administrative code and echocardiogram reports. Thereafter, the features extracted from the ischemic stroke registry of 1,598 patients were used to train 9 different machine learning methods to classify between cardioembolic and non-cardioembolic subtypes. The best performing model was achieved area under the receiver operating characteristic curve of 0.911.

Detecting infarct in non-contrast CT is challenging because of the low contrast or weak signal-to-noise ratio of brain soft tissue (Rekik, Allassonnière et al. 2012). To deal with this problem, certain techniques had been developed. Lee et al. (Lee, Yune et al. 2018) utilized multi-window conversion to increase the conspicuity of certain pathologies. This technique was also applied in our work. Deep convolutional neural network has become state-of-the-art for various medical imaging tasks. Recent studies leveraged the power of deep learning to localize the region of infarct in non-contrast CT. Qiu et al. (Qiu, Kuang et al. 2020) developed an approach to detect early infarction in acute stroke with non-contrast CT. A pretrained segmentation model was used to extract a set of custom-made features on each voxel. Then, the features were processed by a Random Forest to classify the presence of infarct. Pan et al. (Pan, Wu et al. 2021) decomposed each CT slice into a set of different sizes of small patches. Afterward, a convolutional network was used to classify whether each patch is from an infarct area. Moreover, a post-processing method was applied to refine the predicted results. EIS-Net (Kuang, Menon et al. 2021) was a novel method that can simultaneously segment infarct area and provide the Alberta Stroke Program Early CT Score (ASPECTS) (Pexman, Barber et al. 2001). This network contains two major components including Triplet convolutional neural network (T-CNN) for early infarct segmentation and ASPECTS Net for ASPECTS prediction. The first part incorporated context information into the network by comparing symmetric disparity between original CT images, horizontal-flipped images, and the corresponding atlas using comparison disparity blocks (CDB). Then, a multi-level attention gate module (MAGM) was used to fuse the outputs from the CBDs before segmenting the infarct volumes. The second part also utilized MAGM to activate the features relevant to ASPECTS scoring task. Finally, multi-region classification was performed to predict the score of each region.

Despite the novelty, none of these studies incorporate the clinical information and CT images. Though, medical imaging information is normally recorded in EHR, it is still necessary to rely on radiologists to interpret and document the imaging results. To bypass the bottleneck, machine learning techniques could be applied as unified methods to

interpret clinical information together with imaging data. The strategies to fuse medical imaging and EHR were well studied in a work of Huang et al. (Huang, Pareek et al. 2020). They categorized the strategies into three groups which are early, joint, and late fusions. Early fusion integrates features at the input level. Joint fusion joins the hidden features in the middle layers of the model and loss is propagated back to the previous layers. Late fusion aggregates predictions at the output layer. Since we had very small data for cardioembolic stroke prediction but sufficient data for large infarct detection (a patient had multiple CT slices), late fusion was chosen as an appropriate strategy for us. Our large infarct localization model was independently trained with image-level annotations, unlike all the previous infarct detection studies which required pixel or voxel-level annotations on diffusion-weighted imaging (DWI). Finally, the output of the infarct localization model representing the presence of large infarct was aggregated with the clinical features to predict the risk of cardioembolic stroke.

3. Background

3.1.Overview of our approach

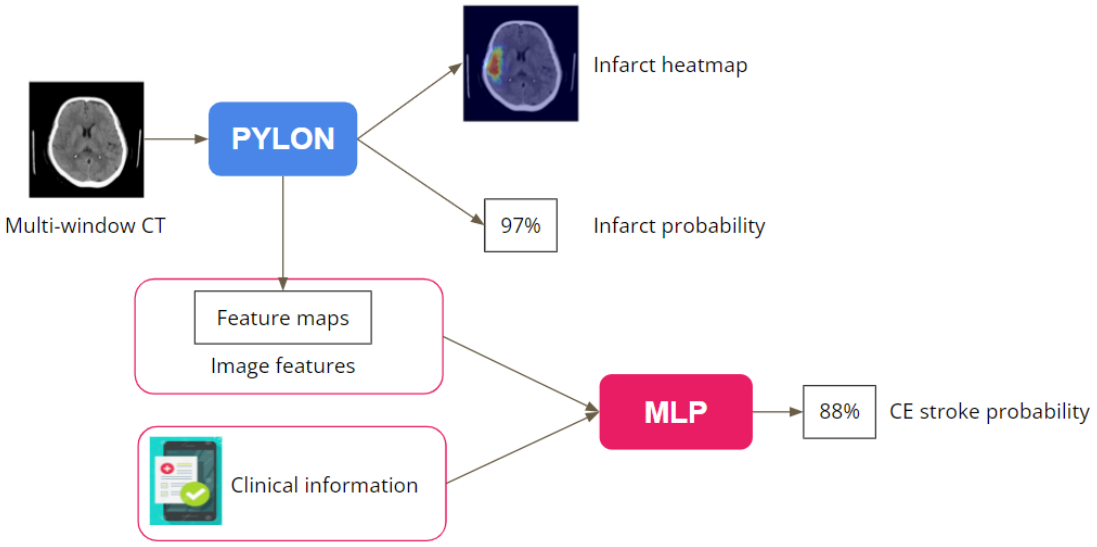


Figure 1: Overview of our end-to-end approach. PYLON is used to extract image features and predict infarct area together with the corresponding infarct probability. The extracted image features and clinical information are then fed into a multi-layer perceptron module to predict the risk of cardioembolic stroke.

Our approach consists of two main modules including an image feature extractor and a joint classifier. We used PYLON as an image feature extractor which can simultaneously localize infarct region and predict the corresponding infarct probability. A multi-layer perceptron module is used as a joint classifier to predict the risk of cardioembolic stroke from the extracted image features and the clinical information. All modules are trained in an end-to-end manner. The overview of our approach is illustrated in Figure 1.

3.2.Neural network

In computer science, neural network, also known as artificial neural network (ANN), is a collection of connected node layers which can transform the input information into intermediate features that are specific to a target. Neural network can learn to extract

features by using backpropagation (Rumelhart, Hinton et al. 1985). There are numerous variants of the neural network nowadays, but the important ones related to our work are described in the following sections.

3.2.1. Multi-Layer Perceptron (MLP)

Multi-layer perceptron (MLP) is a simple class of neural network. It comprises multiple fully connected node layers which transfer the information in a feedforward way. There are three basic types of node layers in MLP including input layer, hidden layer, and output layer. MLP makes use of non-linear activation functions in hidden and output layers to transform the linear combination outcomes into non-linear spaces to distinguish the data that is not linearly separable.

3.2.2. Convolutional Neural Network (CNN)

Convolutional neural network (CNN) is a variant of artificial neural network which is commonly used to extract local pattern in sequential data, e.g. time-series data and image data (LeCun and Bengio 1995). The layer in CNN is called convolutional layer which consists of multiple filters. Each filter is responsible for the extraction of a local pattern which is relatively small compared to the input of the layer. The filters are applied through the input to achieve translation-invariance property. These characteristics of CNN make it much more efficient in local feature extraction compared MLP which must connect every single point of input data to each node in the layer. Therefore, CNN can be considered as a regularized version of MLP because of the immense reduction in network parameters.

3.2.3. Pyramid Localization Network (PYLON)

Pyramid localization network (PYLON) (Preechakul, Sriswasdi et al. 2020) is a specialized convolutional neural network architecture that aims to improve the accuracy of the heatmaps that explain image classification model. The heatmaps can provide accurate localization with only image-level annotation. It uses weakly-supervised learning technique to simultaneously produce both heatmaps and the corresponding classification

scores. The heatmaps are directly derived from the output of the last convolutional layer of the network which contains both fine-grained textural features and context features from the specialized pyramid architecture. The classification scores are simply obtained by applying global max pooling operation on the heatmaps.

3.3. Activation function

The relationship between the input and output data can be either linear or non-linear. However, the node operations in common neural networks are made of basic linear combination. Therefore, to capture non-linear relationship, activation functions were invented to transform the linear outcomes into non-linear spaces. There are two important activation functions used in our work.

3.3.1. Sigmoid

Sigmoid or logistic activation function is a popular activation function used to non-linearly transform a real number into a new domain with a range between 0 and 1. However, this activation function could result in a saturation issue, saturating gradient (Glorot and Bengio 2010). Although this function is not suitable for hidden layers, it is perfect for the output layer when the network is expected to predict likelihoods. The formula of sigmoid function can be written as follows.

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

3.3.2. Rectified Linear Unit (ReLu)

Rectified linear unit (ReLu) is one of the most wildly adopted non-linear activation function because of its simplicity. This activation function transforms an input number into a non-negative number. If the input number is negative, the corresponding output will be zero. In contrast, it will return identical number if the input number is non-negative. Therefore, it can produce non-linear transformation with minor cost and without saturation issue. The formula of ReLu can be written as follows.

$$ReLu\ (x) = \max(0, x) \quad (2)$$

3.4.Cost function

Cost function or objective function is a crucial part for the training of machine learning model. Minimizing the value of cost functions is the goal of training. The selection of the appropriate cost functions depends on the target task. For binary classification, binary cross-entropy (BCE) is a commonly used cost function.

3.4.1. Binary Cross-Entropy (BCE)

Binary cross entropy (BCE) is commonly used in binary classification. When the target is 1, minimizing BCE will result in maximizing the output of the model. When the target is 0, minimizing BCE will result in minimizing the model output. The formula of BCE can be written as follows.

$$BCE(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad (3)$$

Where y is the target label, $\hat{y}$ is the model prediction. Sigmoid function is used as the activation function to convert model output to probability.

3.5.Metric

Metric is used to evaluate the performance of machine learning model. The right metric can indicate the strength and the weakness of the model. Hence, it plays a key role for determining the direction of the model development. The metrics used in this work are described as follows.

3.5.1. Confusion matrix

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

Figure 2: Confusion matrix for binary classification. There are 4 categories of the outcomes including true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

In binary classification, the target label can be either positive or negative while the prediction can also be either positive or negative. Therefore, there are 4 possible categories of outcomes as demonstrated in Figure 2. We can count the total number of the outcomes in each category, and then put it in the corresponding cell in the confusion matrix. When the confusion matrix is completed, we can analyze the outcomes in various aspects. The fundamental metrics are described in the following sections.

3.5.2. Sensitivity

Sensitivity or positive recall is used to determine how well the prediction can recognize the positive samples. It can be calculated as follows.

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

3.5.3. Specificity

Specificity or negative recall is used to determine how well the prediction can recognize the negative samples. It can be calculated as follows.

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

3.5.4. Positive Predictive Value (PPV)

Positive predictive value or positive precision is used to evaluate how accurate the positive prediction is. It can be calculated as follows.

$$PPV = \frac{TP}{TP + FP} \quad (6)$$

3.5.5. Negative Predictive Value (NPV)

Negative predictive value or negative precision is used to evaluate how accurate the negative prediction is. It can be calculated as follows.

$$NPV = \frac{TN}{TN + FN} \quad (7)$$

3.5.6. F1-score

F1-score is a harmonic mean of precision and recall. It is used to evaluate the overall performance of classification with the balance between precision and recall. It can be calculated as follows.

$$F1 - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

3.5.7. Area Under the Curve of Receiver Operating Characteristic (ROC-AUC)

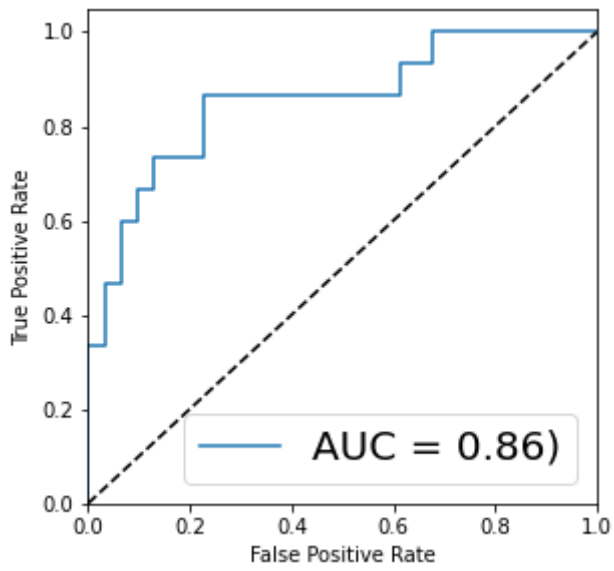


Figure 3: An example of receiver operating characteristic curve (ROC).

Receiver operating characteristic curve (ROC) shows the tradeoff between sensitivity and specificity at different confidence thresholds. As shown in Figure 3, the horizontal axis

represents the degree of false positive rate (FPR) which is equal to 1 – specificity, so the lower value of false positive rate indicates better specificity. The vertical axis represents the scale of true positive rate (TPR) or sensitivity. Since both TPR and FPR are normalized values, the ROC is unbiased to the population sizes. Also, the area under the receiver operating characteristic curve (ROC-AUC) can be used as a performance matric. The higher the ROC-AUC, the better the performance.

3.5.8. Average Precision (AP)

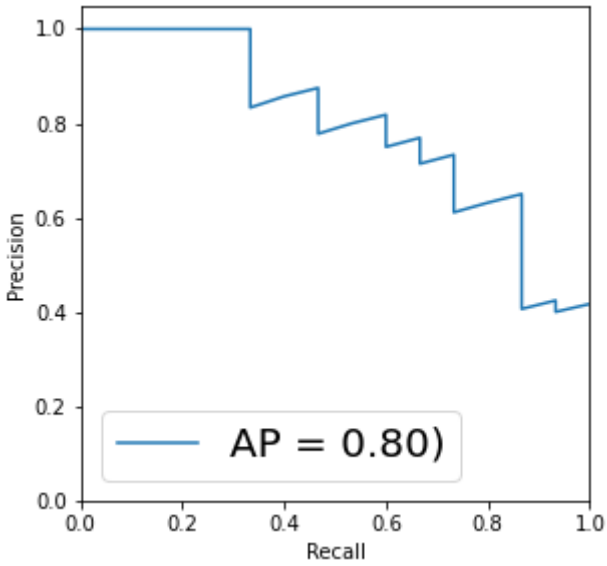


Figure 4: An example of precision-recall curve.

Average precision (AP) is calculated by averaging the precisions at different confidence thresholds. Also, it can be calculated using the area under precision-recall curve as shown in Figure 4. AP can be used as an evaluation matric that summarizes the balance between precision and recall. However, this matric is biased toward the positive population. If positive population are heavily dominant, false positive will be trivial. Consequently, precision can be very high even when all negative samples are missed.

3.6. Shapley additive explanations (SHAP)

Figure 5: A toy example showing the feature impacts (SHAP values) on the model output.

Shapley additive explanations (SHAP) (Lundberg and Lee 2017) is a unified approach used to interpret the predictions of machine learning models. It utilizes game theory to calculate the impacts of the input features toward the model outputs as illustrated in Figure 5. It can provide both scales and directions of the input impacts. Also, SHAP can be applied to any machine learning model, which makes it a powerful tool for model interpretation. We demonstrated the use of SHAP in our experimental results.

4. Data description

The data collection of this study was approved by the Institutional Review Board (IRB) of the Faculty of Medicine, Chulalongkorn University. Two types of the dataset were gathered.

4.1.CT images

The first is CT images in DICOM format of 651 stroke patients at King Chulalongkorn Memorial Hospital (KCMH) collected from 2014 to 2020. Since a patient can have more than a series of CT scans, the total number of CT series is 1,217, comprising 60,334 CT images. Each CT image was labeled by experienced neurologists to identify the presence of a large infarct. A total of 5,840 images were labeled as having large infarcts. However, only image-level annotations were obtained due to limited resources. Therefore, the area of the infarct is undetermined.

4.2.Clinical information

The other dataset is a set of clinical information manually extracted from the EHR of KCMH. According to the complexity of the manual information extraction, only the clinical data of 227 stroke patients from 2019 to 2020 were collected. The CT images of these 227 patients are also included in the aforementioned CT dataset. The clinical information is composed of 49 stroke-relevant features as shown in Table 8 in Appendix A. To mitigate the curse of dimensionality, these features were deliberately refined into 11 important features by experienced neurologists as shown in Table 1. These expert-guided features outperform the raw features in our experiments. The Trial of Org10172 in Acute Stroke (TOAST) is used as the classification label whether the patients had a cardioembolic stroke. TOAST includes 5 subtypes of ischemic stroke: large-artery atherosclerosis (LAA), cardioembolism, small-vessel occlusion, a stroke of other determined etiology, and stroke of undetermined etiology. A patient with the TOAST of cardioembolism is a target of our study. On the other hand, it is unclear whether the patients with undetermined etiology are

considered having a cardioembolic stroke. Therefore, we decided to exclude all patients with undetermined TOAST from our study. The distribution of TOAST is shown in Table 2.

Table 1: The expert-guided features converted from the stroke-relevant features

	Description	Total samples, N=227	Non-CE stroke, N=154	CE stroke, N=73
Demographics				
Female	The patient is female	97 (42.7%)	55 (35.7%)	42 (57.5%)
Age	Age of the patient	65.8 ± 14.3	63.2 ± 13.9	71.4 ± 13.7
Expert-guided features				
Duration LSN	Duration in hours from clear onset or last seen normal to CT	23.7 ± 36.7	28.7 ± 38.3	13.2 ± 30.6
Duration FSA	Duration in hours from clear onset or first seen abnormal to CT	21.5 ± 35.9	26.4 ± 37.4	11.1 ± 30.4
Gradual onset	The patient had stepwise or gradual worsening stroke symptoms	11 (4.8%)	9 (5.8%)	2 (2.7%)
Peak clear onset	The patient had peak stroke symptoms at onset	107 (47.1%)	78 (50.6%)	29 (39.7%)

Cortical lobe sign	The patient had "Dysphasia/Aphasia" or NIHSS 1b = 2 (Answers neither question correctly) or NIHSS 2 = 2 (Forced eye deviation) or NIHSS 3 > 0 (Hemianopia) or NIHSS 9 > 0 (Aphasia) or NIHSS 11 > 0 (Visual, tactile)	84 (37.0%)	33 (21.4%)	51 (69.9%)
Valvular heart disease	The patient had "Valvular heart disease"	11 (4.8%)	1 (0.6%)	10 (13.7%)
Metabolic syndrome	The patient had "Diabetes mellitus" or "Hypertension" or "Obesity" or "Dyslipidemia"	169 (74.4%)	120 (77.9%)	49 (67.1%)
Vascular heart disease	The patient had "Peripheral arterial disease" or "Previous transient ischemic attack " or "Previous stroke" or "Coronary heart disease"	81 (35.7%)	52 (33.8%)	29 (39.7%)
TIA same site	The patient had transient ischemic attack at the same site within 2 weeks	1 (0.4%)	1 (0.6%)	0 (0%)
Smoking	The patient had been smoking	24 (10.6%)	21 (13.6%)	3 (4.1%)

Data displayed as N (%) or mean ± SD. CE denotes cardioembolic. Age is included as a feature.

Table 2: The distribution of TOAST in the clinical data

TOAST	Total samples, N=227	Development set, N=181	Test set, N=46
Large-artery atherosclerosis	55 (24.2%)	46 (25.4%)	9 (19.6%)
Cardioembolism	73 (32.2%)	58 (32%)	15 (32.6%)
Small-vessel occlusion	96 (42.3%)	75 (41.4%)	21 (45.7%)
Other determined	3 (1.3%)	2 (1.1%)	1 (2.2%)

5. Proposed method

5.1.Large infarct detection with PYLON

Infarct detection is crucial in ischemic stroke diagnosis. The presence of a large infarct indicates that there is a blockage in a large blood vessel which can be mainly caused by either LAA or cardioembolism. To detect the large infarct in CT images, we used PYLON as an image classification model to predict the likelihood of large infarct together with its approximate region. The advantage of PYLON is its ability to precisely locate the region of interest, the region of large infarct in this study, which is important for interpretability. All CT images were extracted using multi-window conversion to expose abnormalities in different ranges of density. Each window was obtained by CT windowing, which is commonly used during clinical interpretation by adjusting CT's window-width (WW) and window-length (WL). Three windows were used in this study including brain window (WW=80, WL=40), tissue window (WW=40, WL=40), and blood window (WW=40, WL=60). These windows are demonstrated in Figure 6.

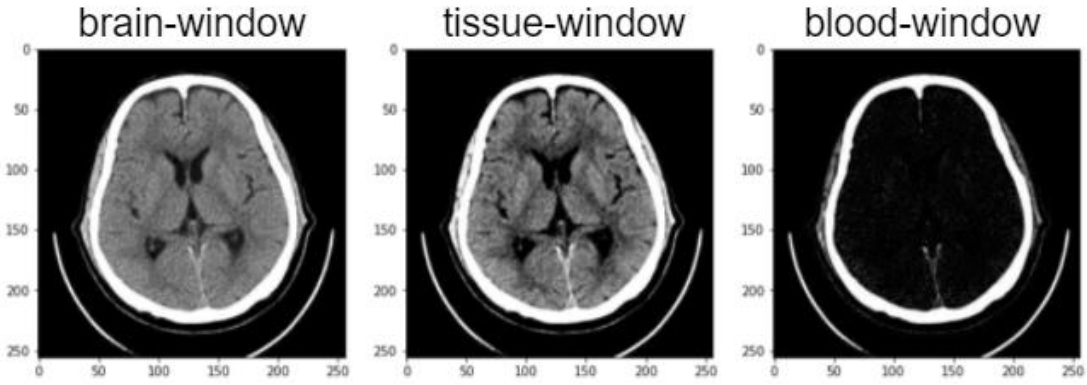


Figure 6: Examples of CT windowing including the brain-window (WW=80, WL=40) for common investigation, tissue-window (WW=40, WL=40) for detecting anomalies in soft tissue, and blood-window (WW=40, WL=60) for detecting high density anomalies such as blood clots.

For each CT, the 3 windows were converted into 8-bit grayscale images and then stacked together as a 3-channel image. The model was trained using these multi-window images as the inputs and infarct labels obtained from the neurologists as the targets. For the model settings, we used the default parameters which have ResNet-50 pre-trained on ImageNet as the encoder. The output dimension of the model was set to 1 for the binary classification of the presence of large infarct. Binary cross entropy (BCE) was used as the objective function for the model training.

We used PyTorch framework to train the model for 100 epochs with the batch size of 64 images. We used Adam to optimize the model's weights with the initial learning rate of 0.0001. The learning rate was conditionally reduced by the factor of 0.1 if the training objective value was not improved after every 10 epochs. The model weights from the epoch with the best validation objective value were chosen for evaluation. We trained our model on a single NVIDIA A100 GPU. To improve the generalizability of our model, we applied several image augmentation techniques including random rotation, random resized crop, horizontal flip, random brightness, and random contrast.

5.2.Cardioembolic stroke prediction

We examined two different approaches for fusing the clinical and CT features. The first approach is late fusion which simply adds the infarct probability obtained from the infarct detection model as an additional feature to the clinical counterpart. Since a CT volume is composed of multiple CT slices, we selected the slice with the maximum infarct probability as the representative of the volume then used its infarct probability as the additional feature. Figure 7A and Figure 7B illustrates the overall architecture of the late fusion approach.

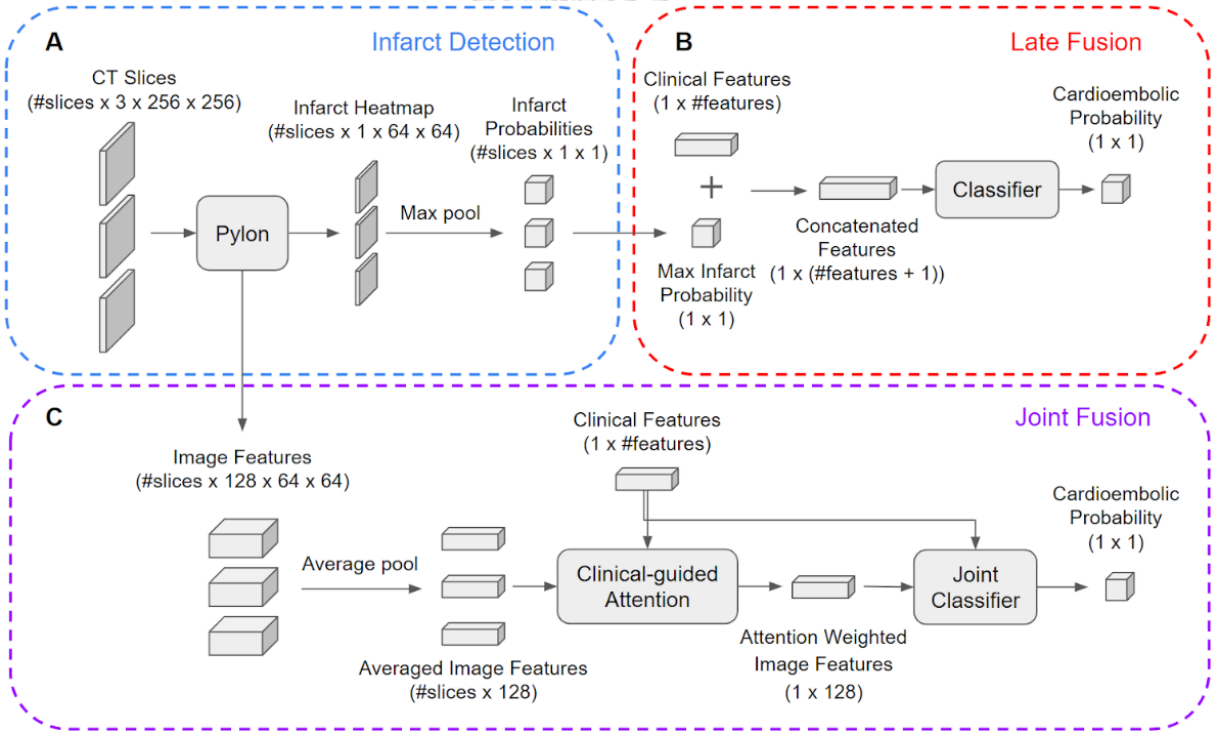


*Figure 7: Overall architectures of our approaches. **A**, infarct detection model producing heatmap for infarct localization. The infarct probability of each CT slice is determined by the maximum value of the corresponding heatmap. **B**, late fusion approach which simply concatenates the max infarct probability with the clinical features. The classifier is independently trained with the concatenated features. **C**, joint fusion approach which can mutually predict infarct probabilities and extract relevant image features for cardioembolic stroke prediction. Joint classifier and clinical-guided attention module are jointly trained with the infarct detection model in an end-to-end manner.*

However, this approach limits the information of CT to the infarct detection outputs which are independent of the stroke subtype classification. Accordingly, we proposed another approach to jointly fuse the CT and clinical information to enable the training of both infarct detection and cardioembolic stroke prediction in an end-to-end manner as shown in Figure 7C. With this approach, the relevant CT features can be directly extracted to suit the stroke classification task. In addition, to handle the variation of the number of slices in an CT series, we developed a clinical-guided attention module to summarize the features of multiple CT slices into one piece. The module assigns appropriate weight for each slice and then outputs the weighted average features. Subsequently, the condensed CT features are concatenated with the clinical features and then fed into a joint classifier to predict the risk of cardioembolic stroke.

5.2.1. Clinical-guided attention

Clinical-guided attention module calculates slice attention weight for each block of the averaged image features. We assumed that the image features from a CT slice can partially imply some clinical information. Therefore, the similarity between the clinical features and the image features can represent the importance of the corresponding CT slice. The block of clinical features is linearly transformed into a query vector in 128-dimensional space. Also, each block of the averaged image features is linearly transformed into a key vector with 128 components. Now, the query vector and the key vectors are in the same vector space. Thus, the similarity (attention activation) between the query vector and the key vector can be represented by the inner product which can be obtained by matrix multiplication. Softmax function is used to normalize the attention activation to the attention weights. Finally, a single block of the attention weighted image features is produced by the matrix multiplication between the blocks of the averaged image features and the attention weights.

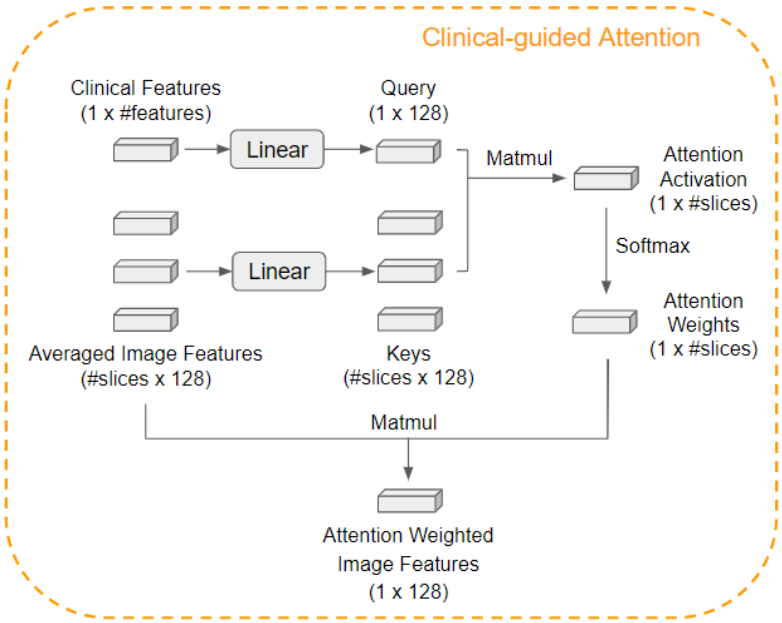


Figure 8: Clinical-guided attention module calculating slice attention weight for each block of the averaged image features.

5.2.2. Joint classifier

Joint classifier module aggregates the clinical features and the image features to predict the risk of cardioembolic stroke. Encoder block is used to non-linearly transform the input vector into a new vector space which is better fit to the downstream task. ReLu is chosen as the non-linear activation function because it is simple, fast and good for gradient descent. Layer normalization is used to stabilize the output of the linear layer resulting in faster convergence during training. We used 3 encoder blocks to consecutively encode the input feature vector. The first, second and third blocks output 128-dimensional, 64-dimensional and 32-dimensional output feature vectors respectively. The encoded image features are concatenated with the clinical features to form a vector of multi-modal features. Then, the feature vector is encoded by 3 encoder blocks. Lastly, the encoded feature vector is non-linearly transformed by a linear layer with sigmoid activation function to produce the probability of cardioembolic stroke.

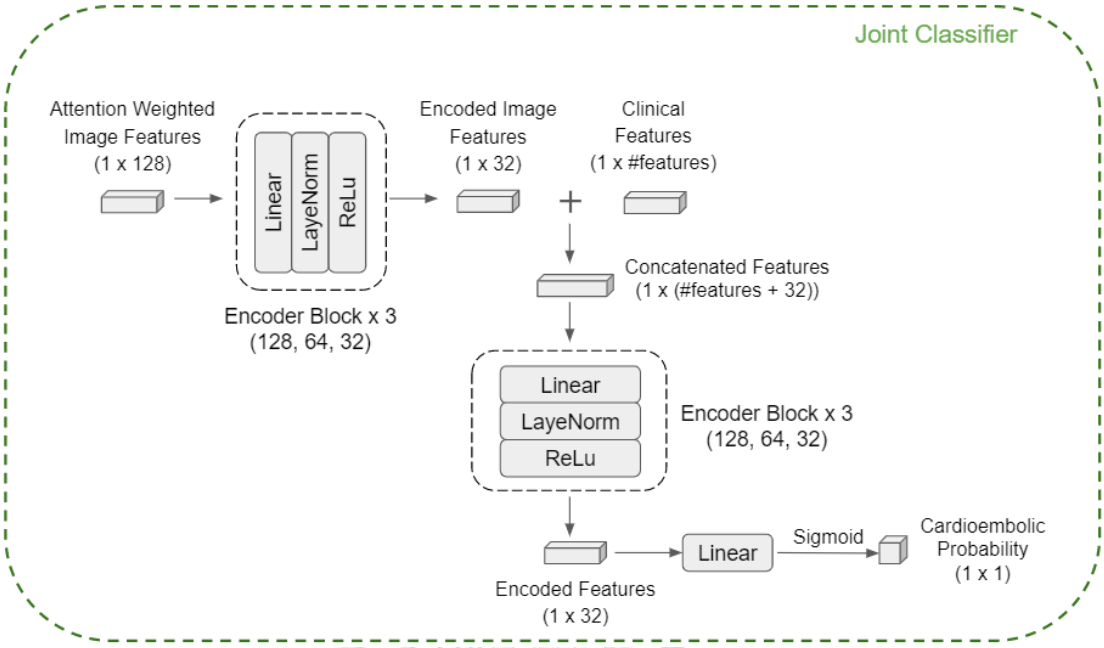


Figure 9: Joint classifier module aggregating the clinical features and the image features to predict the risk of cardioembolic stroke

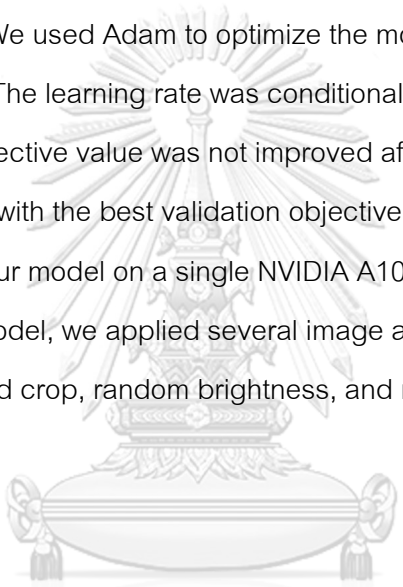
Since the dataset of the cardioembolic stroke prediction task is relatively small, we firstly trained the infarct detection model on the full training dataset of NCCT dataset which to obtain the pre-trained weights for the subsequent training of the joint-fusion model. Then we jointly re-trained the pre-trained infarct detection model along with the clinical-guide attention module and the joint classifier on the development dataset of the cardioembolic stroke prediction task. There are two objective functions in the joint training process. The first objective is to minimize the binary cross entropy (BCE) for the infarct detection task in the same way as the pre-training step. Also, the second objective is to minimize the BCE for the cardioembolic stroke prediction task. These objectives can be written in a unified form as follows.

$$Loss_{joint} = \lambda_{inf} BCE_{inf} + \lambda_{ce} BCE_{ce} \quad (9)$$

Where $Loss_{joint}$ is the unified objective, λ_{inf} and λ_{ce} are the real number coefficients for the objective of the infarct detection task (BCE_{inf}) and the objective of the

cardioembolic stroke prediction task (BCE_{ce}) respectively. We used $\lambda_{inf} = 1$ and $\lambda_{ce} = 1$ in our study.

PyTorch framework was used to train the joint fusion model for 30 epochs with the batch size of 32 samples. Due to the variation of the number of NCCT slices, each sample data including a single clinical feature vector and a set of multi-windowed NCCT images was individually fed into the joint-fusion model. The gradient of each sample in the batch is accumulated one by one. At the batch end, the model is optimized with the accumulated gradient. We used Adam to optimize the model's weights with the initial learning rate of 0.0001. The learning rate was conditionally reduced by the factor of 0.1 if the unified training objective value was not improved after every 10 epochs. The model weights from the epoch with the best validation objective value were chosen for evaluation. We trained our model on a single NVIDIA A100 GPU. To improve the generalizability of our model, we applied several image augmentation techniques including random resized crop, random brightness, and random contrast.



6. Experimental results

6.1. Experiment setup

We assessed 7 common machine learning classifiers for distinguishing between cardioembolic and non-cardioembolic strokes following the late fusion approach, including K-Nearest Neighbors (KNN), Logistic Regression (Logistic), Support Vector Machine (SVM), Decision Tree (Tree), Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and Multi-Layer Perceptron (MLP). For the joint fusion approach, we only assessed the performance of the proposed joint fusion model.

We defined the cardioembolic risk estimation of stroke patients as a binary classification problem. The patients with the TOAST classification of cardioembolism were labeled as the target or positive samples, while the rest were labeled as negative. This results in a total number of 73 positive samples and 154 negative samples. These samples were randomly split into a development set and a test set of 184 and 46 samples respectively. Then, the development set was split into train set and validation set using stratified k-fold with k equal to 5. The split of these samples is shown in Table 3. For the CT image dataset, the images of the 227 patients were split by patient's ID to match the split of the previous dataset. Also, the images of the remaining 424 patients were split by patient's ID using the same strategy as mentioned earlier. The splitting label for the stratified k-fold of the remaining patients was the indicator of whether the patient has an image with a large infarct label. Table 4 illustrates the distribution of the numbers of images in each split.

Table 3: The number of samples (and the positive samples) in each split of the clinical information for the cardioembolic risk estimation task.

Set	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	144 (46)	145 (47)	145 (47)	145 (46)	145 (46)
Validation	37 (12)	36 (11)	36 (11)	36 (12)	36 (12)
Test	46 (15)	46 (15)	46 (15)	46 (15)	46 (15)

Table 4: The number of images (and the images with large infarct) in each split of the CT dataset for the large infarct detection task.

Set	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Train	37,965 (4,001)	38,391 (3,597)	38,636 (3,695)	39,389 (3,906)	39,067 (3,697)
Validation	10,397 (723)	9,971 (1,127)	9,726 (1,029)	8,973 (818)	9,295 (1,027)
Test	11,972 (1,116)	11,972 (1,116)	11,972 (1,116)	11,972 (1,116)	11,972 (1,116)

6.2.Performance comparison

6.2.1. Large infarct detection in CT image

Our approach achieved high performance in the large infarct detection task, image-level binary classification in this context, with the average area under the receiver operating characteristic curve (ROC-AUC) over 90 percent. Also, we compare the performance of the model trained with multi-window CT images to that of the model trained with only regular brain-window CT images. The overall performance of the latter was slightly inferior as shown in Table 5 and Figure 10.

Table 5: The comparison of ROC-AUC on the test set of infarct detection task

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVG.	STD.
Brain-window	0.9102	**0.9147**	0.9058	0.9152	0.8936	0.9079	**0.0079**
Multi-window	**0.9184**	0.8839	**0.9168**	**0.9160**	**0.9115**	**0.9093**	0.0129

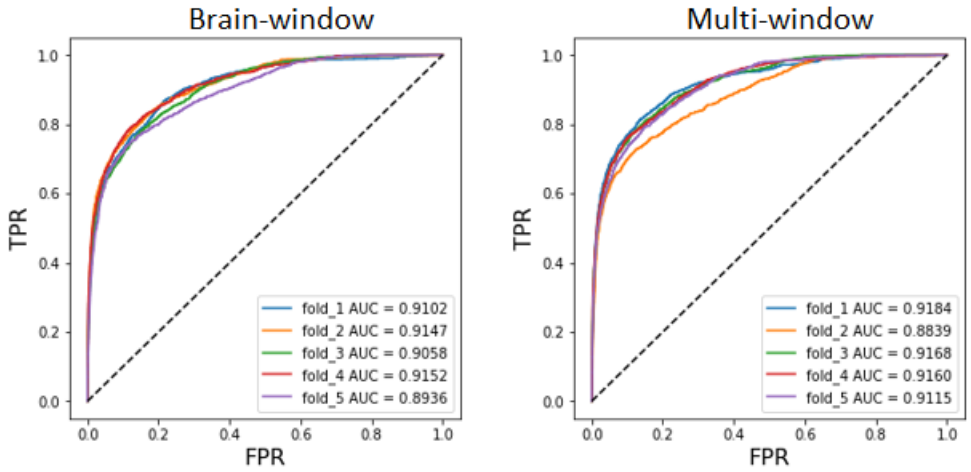


Figure 10: ROC comparison between brain-window and multi-window.

In contrast, the average precisions of these two approaches are comparable. Although multi-window performed better in 3 folds, its average performance is slightly lower than that of the brain-window as illustrated in Table 6 and Figure 11.

Table 6: The comparison of average precision on the test set of infarct detection task

Approach	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	AVG.	STD.
Brain-window	0.6552	**0.6660**	0.6396	**0.6751**	0.6126	**0.6497**	**0.0220**
Multi-window	**0.6725**	0.5961	**0.6669**	0.6636	**0.6319**	0.6462	0.0287

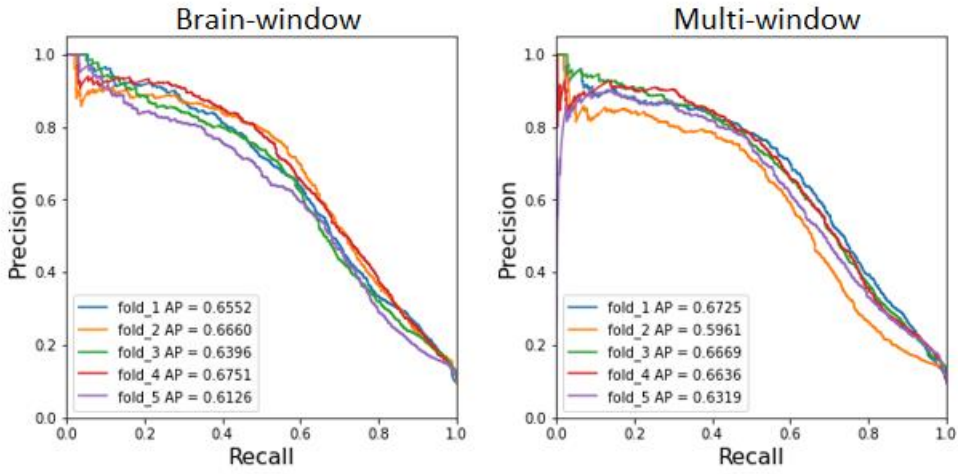


Figure 11: PRC comparison between brain-window and multi-window.

6.2.2. Cardioembolic stroke prediction

By using the infarct probability in addition to the clinical features for cardioembolic stroke prediction task, the performances of all classifiers improve as demonstrated in Table 7. Among the 7 late fusion classifiers, SVM outperformed the others with the average ROC-AUC of 78.8% when using full clinical features, and 81.8% when using the expert-guided features. K-nearest neighbors, random forest and multi-layer perceptron with expert-guided features also performed relatively well with the mean ROC-AUC greater than 78%. On the other hand, the joint fusion method with the clinical-guided attention module achieved the best performance with the average ROC-AUC of 84.0%.

Table 7: The comparison of ROC-AUC on the test set of the stroke prediction task.

Approach	Tree	XGBoost	Logistic	RF	KNN	MLP	SVM	Joint Fusion
Full clinic	0.64 (0.05)	0.69 (0.03)	0.71 (0.04)	0.75 (0.05)	0.64 (0.05)	0.71 (0.05)	0.78 (0.03)	
Expert clinic	0.68 (0.02)	0.74 (0.03)	0.77 (0.01)	0.77 (0.02)	0.75 (0.06)	0.77 (0.03)	0.80 (0.07)	
Full clinic + CT	**0.74 (0.05)**	0.73 (0.05)	0.71 (0.03)	0.77 (0.02)	0.66 (0.09)	0.73 (0.05)	0.79 (0.02)	0.72 (0.07)
Expert clinic + CT	0.72 (0.02)	**0.75 (0.04)**	**0.77 (0.01)**	**0.78 (0.03)**	**0.78 (0.03)**	**0.80 (0.02)**	**0.82 (0.06)**	**0.84 (0.02)**

*Data displayed as mean (SD)

6.3. Interpretability

6.3.1. Localization of large infarct in CT image

With PYLON, the region of the large infarct is simultaneously provided together with its probability in a single inference. This is similar to the segmentation problem that requires pixel-level annotations to supervise the training process. Nonetheless, our approach needs only image-level annotation to achieve localization capability. The predicted outcome can be converted into a heatmap as illustrated in Figure 12.

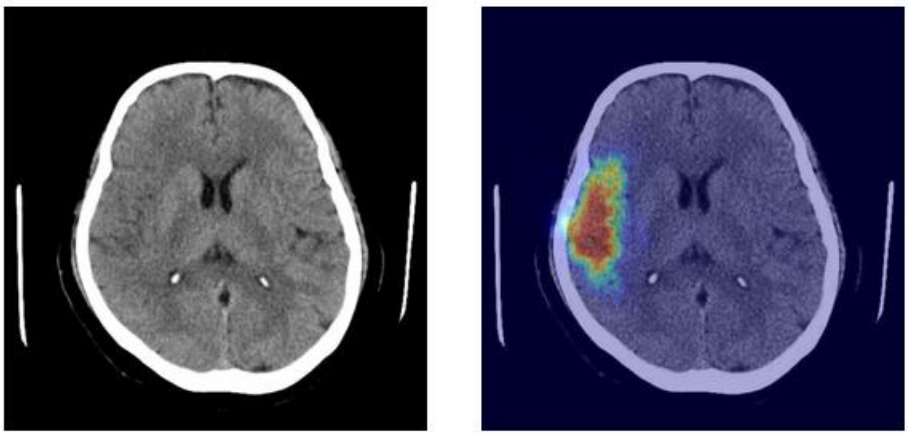


Figure 12: An example of large infarct localization with only image-level annotation. The heatmap color on the right represents the probability that the pixel is part of the large infarct region.

6.3.2. Feature explanation

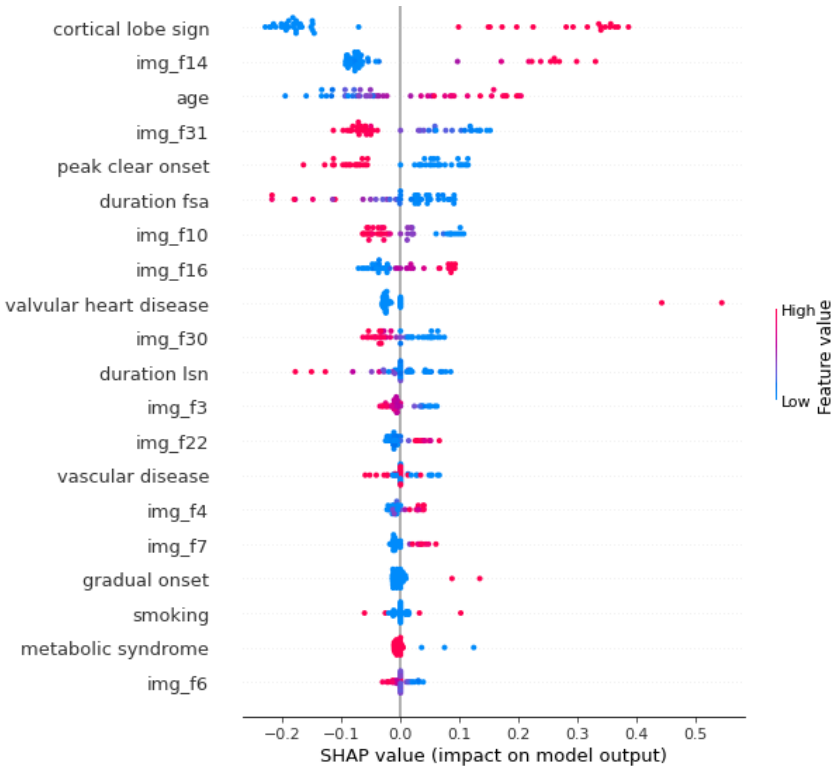


Figure 13: SHAP values indicating the impact of the input features on the joint-fusion model output. Positive impacts increase the value of output probability, while negative impacts reduce the output probability. The most impactful features are sorted from top to bottom. The feature values are displayed in color, blue color indicates low feature value while red color indicates high feature value. Image features are denoted by "img_f".

Trust is a key to enabling the adoption of machine learning in the real world, especially in healthcare. Therefore, it is important to be able to understand the decision of the model. SHAP value is a way to interpret the impact of each input feature on the model output. By applying SHAP to our model, from Figure 13, the most impactful features are cortical lobe

sign, image features, age, onset and valvular heart disease. The impacts of these features are consistent with medical knowledge. The presence of cortical lobe signs increases the risk of cardioembolic stroke. Also, older patients tend to have a higher risk of atrial fibrillation (AF) which is a potential cause of cardioembolic stroke. Moreover, the damage caused by cardioembolic stroke is likely to be severe leading to highly noticeable symptoms. Thus, the people around can quickly detect and call for help. This demonstrates that our model's decision was interpretable and reasonable.

7. Discussion

The results in the infarction detection section were quite counterintuitive because of the negligible effect of multi-window conversion previously perceived as very helpful. This suggests that only brain-window might be sufficient for the task. Nevertheless, we decided to continue to apply multi-window conversion since it often outperformed the brain-window. Another concern with this dataset is the high degree of imbalance. This reflects on the low AP's. We tried several techniques to tackle this problem such as oversampling and Focal loss. Unfortunately, they were not effective.

For cardioembolic stroke prediction, it is obvious that the combination of CT and expert-guided features achieved the best performance. CT alone might not be sufficient to classify ischemic stroke subtypes. Particularly for the patients with early hospital admission, their brain tissue may look normal, so CT might not capture any abnormality. Conversely, clinical information could be analyzed to infer the subtypes of stroke. This could be explained by the fact that the cause of stroke is a brain injury. Thus, the location of the brain damage could be inferred by the characteristic of the stroke symptoms. Adding only infarct probability to the clinical features could provide performance gain to all the proposed classifiers. Furthermore, jointly training the infarct detection task and the cardioembolic stroke prediction task with the clinical-guided attention module yielded even better results. However, it is quite difficult to learn the relationship between the symptoms and brain damage if the data is small. We had only 227 samples while the size of the clinical features is 49. This could introduce the curse of dimensionality and harm the model performance. Fortunately, with expert-guided features whose dimension is only 11, the model performances were significantly improved. Although the integration of CT and expert-guided features were effective, the best performance was still moderate with the best ROC-AUC of 0.86. We found that our model often misclassified large artery atherosclerosis as cardioembolism as demonstrated in Figure 14. This may be due to the

similarity of the clinical and brain imaging findings between them. Therefore, we may need additional information or more amount of data to reach higher performance.

Our study had several limitations. The dataset used in the development of the stroke classification models was relatively small (N=227). This may limit the generalization performance of our machine learning models. Thus, to scale for practical usage, a larger dataset is recommended to improve the model's robustness. Moreover, although the interpretation of the clinical features was comprehensible, the image features extracted from the infarct detection model were obscure. This could be considered as the trade-off between interpretability and performance, using only infarct probability as the image feature was interpretable while using the extracted image features yielded better performance.

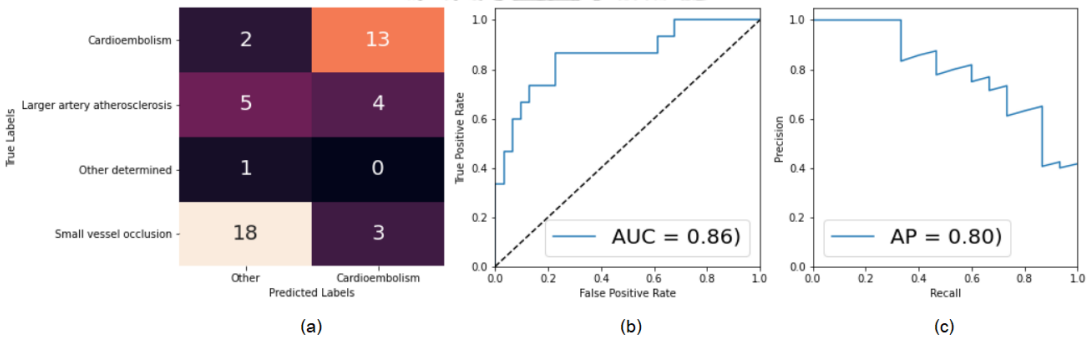


Figure 14: Summary of the performance on the test set in fold 2 using both max CT score and expert-guided features as the input. (a) Confusion matrix divided into 4 classes of TOAST, the prediction threshold was selected based on the optimal sensitivity score on the validation set. The majority of false-positive is large artery atherosclerosis which is similar to cardioembolism in both clinical and brain imaging findings. (b) ROC of the corresponding cardioembolic stroke prediction task with AUC = 0.86. (c) Precision-Recall curve of the task with the average precision of 0.80.

8. Conclusion

8.1. Conclusion

Our study demonstrates how the risk of cardioembolic stroke can be estimated using both clinical information and non-contrast CT images which are normally available in general hospitals. Our approach not only provides the evaluation of the risk of cardioembolic stroke but also provides the interpretability of the model decision in the forms of the heatmap for large infarct localization and the feature impacts for interpretation. The large infarct detection performance achieved the average ROC-AUC of 0.909 with only image-level annotations from experienced neurologists. Even with the small data of 227 samples, our method was able to reach the average ROC-AUC of 0.840 in cardioembolic stroke prediction.

8.2. Future work

Despite the accomplishment of combining the clinical and non-contrast CT features in our current work, the proposed method limits the role of clinical information to the calculation of the slice attention weights. To further exploit the clinical information, it could be used to determine the important regions within CT images to further extract more informative CT features. If this is possible, it may not only benefit the model performance but also the interpretability of the CT features. Therefore, we suggest the extraction of spatial CT features based on the clinical information as a potential direction for the future work.

9. Appendix A. Data description for the 49 stroke-relevant features

Table 8: The stroke-relevant features manually extracted from EHR

	Description	Total samples, N=227	Non-CE stroke, N=154	CE stroke, N=73
Demographics				
Female	The patient is female	97 (42.7%)	55 (35.7%)	42 (57.5%)
Age	Age of the patient	65.8 ± 14.3	63.2 ± 13.9	71.4 ± 13.7
Full stroke-relevant features				
Duration LSN	Duration in hours from clear onset or last seen normal to CT	23.7 ± 36.7	28.7 ± 38.3	13.2 ± 30.6
Duration FSA	Duration in hours from clear onset or first seen abnormal to CT	21.5 ± 35.9	26.4 ± 37.4	11.1 ± 30.4
Wake-up onset	The patient woke up with stroke symptoms	54 (23.8%)	38 (24.7%)	16 (21.9%)
Peak clear onset	The patient suddenly showed stroke symptoms at onset	148 (65.2%)	95 (61.7%)	53 (72.6%)
Gradual onset	The patient had stepwise or gradual worsening stroke symptoms	11 (4.8%)	9 (5.8%)	2 (2.7%)

Rapidly improve	The patient had rapidly improving stroke symptoms	2 (0.9%)	2 (1.3%)	0 (0%)
Heart rate	Heart rate of the patient (BPM)	82.2 ± 16.4	80.5 ± 13.3	85.8 ± 21.3
SBP	Systolic blood pressure (mmHg)	156.1 ± 26.1	157.6 ± 26.5	152.8 ± 25
DBP	Diastolic blood pressure (mmHg)	87.6 ± 16.8	88.3 ± 15.6	86.1 ± 19.3
NIHSS 1a	Level of consciousness	0.2 ± 0.5	0.1 ± 0.3	0.4 ± 0.7
NIHSS 1b	Level of consciousness questions	0.3 ± 0.7	0.1 ± 0.5	0.8 ± 0.9
NIHSS 1c	Level of consciousness commands	0.2 ± 0.6	0.1 ± 0.4	0.5 ± 0.8
NIHSS 2	Best gaze	0.3 ± 0.7	0.1 ± 0.4	0.7 ± 0.9
NIHSS 3	Visual field	0.2 ± 0.6	0.1 ± 0.3	0.5 ± 0.9
NIHSS 4	Facial palsy	0.8 ± 0.9	0.6 ± 0.8	1.1 ± 1
NIHSS 5a	Motor arm (left)	0.8 ± 1.2	0.6 ± 1	1.2 ± 1.5
NIHSS 5b	Motor arm (right)	0.7 ± 1.2	0.5 ± 0.9	1 ± 1.6
NIHSS 6a	Motor leg (left)	0.7 ± 1.2	0.5 ± 0.9	1.3 ± 1.5
NIHSS 6b	Motor leg (right)	0.6 ± 1.1	0.4 ± 0.9	1 ± 1.5

NIHSS 7	Limb ataxia	0.2 ± 0.5	0.2 ± 0.5	0.2 ± 0.4
NIHSS 8	Sensory	0.5 ± 0.6	0.4 ± 0.5	0.6 ± 0.7
NIHSS 9	Best language	0.5 ± 1	0.2 ± 0.7	1 ± 1.3
NIHSS 10	Dysarthria	0.6 ± 0.7	0.5 ± 0.6	0.8 ± 0.7
NIHSS 11	Extinction and inattention	0.2 ± 0.6	0.1 ± 0.4	0.6 ± 0.9
Alteration of consciousness	The patient had alteration of consciousness	22 (9.7%)	6 (3.9%)	16 (21.9%)
Right facial weakness	The patient had right facial weakness	30 (13.2%)	14 (9.1%)	16 (21.9%)
Left facial weakness	The patient had left facial weakness	43 (18.9%)	24 (15.6%)	19 (26%)
Right hemiparesis	The patient had right hemiparesis	80 (35.2%)	52 (33.8%)	28 (38.4%)
Left hemiparesis	The patient had left hemiparesis	87 (38.3%)	55 (35.7%)	32 (43.8%)
Right hypoesthesia	The patient had right hypoesthesia	24 (10.6%)	20 (13%)	4 (5.5%)
Left hypoesthesia	The patient had left hypoesthesia	21 (9.3%)	18 (11.7%)	3 (4.1%)
Dysarthria	The patient had dysarthria	72 (31.7%)	49 (31.8%)	23 (31.5%)
Dysphasia/aphasia	The patient had dysphasia/aphasia	42 (18.5%)	15 (9.7%)	27 (37%)

Ataxia	The patient had ataxia	19 (8.4%)	17 (11%)	2 (2.7%)
Vertigo	The patient had vertigo	13 (5.7%)	11 (7.1%)	2 (2.7%)
Diplopia	The patient had diplopia	2 (0.9%)	1 (0.6%)	1 (1.4%)
Visual problem	The patient had visual problem	11 (4.8%)	5 (3.2%)	6 (8.2%)
TIA same site	Transient ischemic attack (TIA) at the same site within 2 weeks	1 (0.4%)	1 (0.6%)	0 (0%)
Previous TIA	The patient previously had TIA	0 (0%)	0 (0%)	0 (0%)
Previous stroke	The patient previously had stroke	54 (23.8%)	37 (24%)	17 (23.3%)
HT	The patient had hypertension	147 (64.8%)	102 (66.2%)	45 (61.6%)
DM	The patient had diabetes mellitus	78 (34.4%)	58 (37.7%)	20 (27.4%)
DLP	The patient had dyslipidemia	86 (37.9%)	64 (41.6%)	22 (30.1%)
Valvular heart disease	The patient had valvular heart disease	11 (4.8%)	1 (0.6%)	10 (13.7%)
Coronary heart disease	The patient had coronary heart disease	33 (14.5%)	18 (11.7%)	15 (20.5%)
CKD	The patient had chronic kidney disease	16 (7%)	7 (4.5%)	9 (12.3%)

Peripheral arterial disease	The patient had peripheral arterial disease	2 (0.9%)	1 (0.6%)	1 (1.4%)
Obesity	The patient had obesity	1 (0.4%)	1 (0.6%)	0 (0%)
Smoking	The patient had been smoking	24 (10.6%)	21 (13.6%)	3 (4.1%)
Malignancy	The patient had malignancy	13 (5.7%)	6 (3.9%)	7 (9.6%)

Data displayed as N (%) or mean ± SD. CE denotes cardioembolic. Age is included as a feature.

REFERENCES

Amarenco, P., et al. (2009). "Classification of Stroke Subtypes." Cerebrovascular Diseases **27**: 493-501.

Amarenco, P., et al. (2013). "The ASCOD phenotyping of ischemic stroke (Updated ASCO Phenotyping)." Cerebrovascular diseases (Basel, Switzerland) **36**: 1-5.

Amort, M., et al. (2012). "Etiological Classifications of Transient Ischemic Attacks: Subtype Classification by TOAST, CCS and ASCO – A Pilot Study." Cerebrovascular Diseases **33**: 508-516.

Ay, H., et al. (2007). "A computerized algorithm for etiologic classification of ischemic stroke: The causative classification of stroke system." Stroke **38**: 2979-2984.

Bamford, J., et al. (1991). "Classification and natural history of clinically identifiable subtypes of cerebral infarction." Lancet (London, England) **337**: 1521-1526.

Fang, G., et al. (2020). "Automated ischemic stroke subtyping based on machine learning approach." IEEE Access **8**: 118426-118432.

Garg, R., et al. (2019). "Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing." Journal of Stroke and Cerebrovascular Diseases **28**: 2045-2051.

Glorot, X. and Y. Bengio (2010). Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings.

Guan, W., et al. (2020). "Automated Electronic Phenotyping of Cardioembolic Stroke." Stroke: 181-189.

HP, A., et al. (1993). "Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment." Stroke **24**: 35-41.

Huang, S. C., et al. (2020). "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines." npj Digital Medicine 2020 3:1 **3**: 1-9.

Kuang, H., et al. (2021). "EIS-Net: Segmenting early infarct and scoring ASPECTS simultaneously on non-contrast CT of patients with acute ischemic stroke." Medical Image Analysis **70**.

LeCun, Y. and Y. Bengio (1995). "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks **3361**(10): 1995.

Lee, H., et al. (2018). "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets." Nature Biomedical Engineering 2018 3:3 **3**: 173-182.

Lundberg, S. M. and S.-I. Lee (2017). "A unified approach to interpreting model predictions." Advances in neural information processing systems **30**.

Pan, J., et al. (2021). "Detecting the Early Infarct Core on Non-Contrast CT Images with a Deep Learning Residual Network." Journal of Stroke and Cerebrovascular Diseases **30**.

Pexman, J. H. W., et al. (2001). "Use of the Alberta Stroke Program Early CT Score (ASPECTS) for Assessing CT Scans in Patients with Acute Stroke." AJNR: American Journal of Neuroradiology **22**: 1534.

Preechakul, K., et al. (2020). "High resolution weakly supervised localization architectures for medical images."

Qiu, W., et al. (2020). "Machine Learning for Detecting Early Infarction in Acute Stroke with Non–Contrast-enhanced CT." Radiology **294**: 638-644.

Rekik, I., et al. (2012). "Medical image analysis methods in MR/CT-imaged acute-subacute ischemic stroke lesion: Segmentation, prediction and insights into dynamic evolution simulation models. A critical appraisal." NeuroImage: Clinical **1**: 164-178.

Rumelhart, D. E., et al. (1985). Learning internal representations by error propagation, California Univ San Diego La Jolla Inst for Cognitive Science.

Sung, S. F., et al. (2020). "EMR-Based Phenotyping of Ischemic Stroke Using Supervised Machine Learning and Text Mining Techniques." IEEE Journal of Biomedical and Health Informatics **24**: 2922-2931.

VITA

NAME	Pasit Jakkrawankul
DATE OF BIRTH	5 January 1995
PLACE OF BIRTH	Bangkok
INSTITUTIONS ATTENDED	Chulalongkorn University
HOME ADDRESS	20/53 Passorn13 Village, Suwinthawong Road, Lam Phak Chi, Nong Chok, Bangkok 10530