

DEPRESSION CLASSIFICATION ON PRIVACY PROTECTED FACIAL FEATURES DATA



A Thesis Submitted in Partial Fulfillment of the Requirements  
for the Degree of Master of Engineering in Computer Engineering  
Department of Computer Engineering  
Faculty Of Engineering  
Chulalongkorn University  
Academic Year 2023

การจำแนกภาวะซึมเศร้าจากข้อมูลลักษณะใบหน้าที่ได้รับการคุ้มครองความเป็นส่วนตัว



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต  
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์  
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย  
ปีการศึกษา 2566

Thesis Title                                   DEPRESSION CLASSIFICATION ON PRIVACY PROTECTED  
FACIAL FEATURES DATA  
By   Miss Yanisa Mahayossanunt  
Field of Study                                 Computer Engineering  
Thesis Advisor                               Associate Professor PEERAPON VATEEKUL, Ph.D.  
Thesis Co Advisor                           Associate Professor SOLAPHAT HEMRUNGROJN, M.D.

---

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

..... Dean of the FACULTY OF  
ENGINEERING  
(Professor SUPOT TEACHAVORASINSKUN, D.Eng.)

THESIS COMMITTEE

..... Chairman  
(Assistant Professor NATAWUT NUPAIROJ, Ph.D.)  
..... Thesis Advisor  
(Associate Professor PEERAPON VATEEKUL, Ph.D.)  
..... Thesis Co-Advisor  
(Associate Professor SOLAPHAT HEMRUNGROJN, M.D.)  
..... Examiner  
(PUNNARAI SIRICHAROEN, Ph.D.)  
..... External Examiner  
(Assistant Professor Thanapat Kangkachit, Ph.D.)

ญาณิศา มหายศนันท์ : การจำแนกภาวะซึมเศร้าจากข้อมูลลักษณะใบหน้าที่ได้รับการ  
 คัดกรองความเป็นส่วนตัว. ( DEPRESSION CLASSIFICATION ON PRIVACY  
 PROTECTED FACIAL FEATURES DATA) อ.ที่ปรึกษาหลัก : รศ. ดร.พีรพล เวทีกุล, อ.  
 ที่ปรึกษาร่วม : รศ. พญ.โสฬสพัทธ์ เหมรัญชโรจน์

วิทยานิพนธ์ฉบับนี้นำเสนอการจำแนกภาวะซึมเศร้าจากข้อมูลลักษณะใบหน้าที่ได้รับ  
 ความคุ้มครองความเป็นส่วนตัว การจำแนกโรคซึมเศร้าได้อย่างรวดเร็วเพื่อให้ผู้ป่วยได้รับการรักษา  
 อย่างทัน่วงที่เป็นวิธีการที่สามารถป้องกันความเสียหายจากโรคได้ อย่างไรก็ตามการจำแนกโรค  
 ซึมเศร้าได้อย่างรวดเร็วและมีประสิทธิภาพเป็นความท้าทายอย่างยิ่ง เนื่องจากบุคคลากรทางการแพทย์  
 แพทย์ที่ไม่เพียงพอและระยะเวลาในการวินิจฉัยโรคนั้นใช้เวลานานต่อผู้ป่วยหนึ่งคน การนำ  
 ปัญญาประดิษฐ์มาช่วยในการจำแนกโรคจึงมีประโยชน์ในการช่วยลดภาระของแพทย์ แต่การนำ  
 ปัญญาประดิษฐ์มาใช้ในทางการแพทย์ก็มีความท้าทายในด้านการปกป้องความเป็นส่วนตัวของ  
 ผู้ป่วย ในที่นี้เราจึงได้ใช้ข้อมูลลักษณะใบหน้าที่ได้รับการสกัดมาจากสีหน้าของผู้ป่วยขณะทำการ  
 สัมภาษณ์ทางจิตเวชมาเป็นข้อมูลในการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง แบบจำลองมีการ  
 ใช้ เทคนิค LSTM, Attention Mechanism, Intermediate Fusion, และ Label  
 Smoothing ในการเพิ่มประสิทธิภาพของแบบจำลอง งานวิจัยนี้ได้ดำเนินการบนวิธีทัศน์บท  
 สัมภาษณ์ 474 วิธีทัศน์ซึ่งถูกรวบรวมโดยจุฬาลงกรณ์มหาวิทยาลัย แบ่งเป็นวิธีทัศน์ ผู้ป่วยโรค  
 ซึมเศร้า 134 วิธีทัศน์ และผู้ไม่ป่วยโรคซึมเศร้า 340 วิธีทัศน์ จากการทดสอบพบว่าแบบจำลอง  
 สามารถทำคะแนนประสิทธิภาพได้ดังต่อไปนี้ 91.67% accuracy, 91.40% precision, 87.03%  
 recall, และ 88.89% F1-score นอกจากนี้แบบจำลองยังถูกนำไปวิเคราะห์ด้วยวิธีการ  
 Integrated Gradient ซึ่งสามารถอธิบายความสัมพันธ์ระหว่างข้อมูลลักษณะใบหน้าที่เกี่ยวข้องกับ  
 โรคซึมเศร้าได้ จากผลการวิเคราะห์แบบจำลองข้อมูลลักษณะใบหน้าที่มีความสำคัญในการจำแนก  
 โรคซึมเศร้าคือ ผู้ป่วยที่มีอาการโรคซึมเศร้าจะหันหน้าหนีกล้อง มีดวงตาเลื่อนลอย กวาดสายตา  
 ได้ช้า ไม่ยิ้ม ขมวดคิ้ว และทำหน้าตาบูดบึ้ง ซึ่งแสดงถึงอาการไม่มีสมาธิ การปลีกตัวออกจาก  
 สังคม และความรู้สึกเชิงลบซึ่งเป็นอาการตามปกติของโรคซึมเศร้า

สาขาวิชา วิศวกรรมคอมพิวเตอร์

ลายมือชื่อนิสิต .....

ปีการศึกษา 2566

ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

# # 6470165021 : MAJOR COMPUTER ENGINEERING

KEYWORD: depression detection, facial expression, deep learning

Yanisa Mahayossanunt : DEPRESSION CLASSIFICATION ON PRIVACY PROTECTED FACIAL FEATURES DATA. Advisor: Assoc. Prof. PEERAPON VATEEKUL, Ph.D. Co-advisor: Assoc. Prof. SOLAPHAT HEMRUNGROJN, M.D.

This thesis presents depression classification on privacy protected facial features data. Fast depression classification to help patients receive proper treatment is a method that can prevent the damage of depression. However, fast and effective depression classification is difficult because medical personnel are adequate and the time to analyze depression is long per patient. Applied artificial intelligence in the medical field can help reduce the workload of medical personnel. It is also difficult because of privacy protection. Therefore, we utilize extracted facial features from facial expressions in clinical interview videos to develop a machine learning model. The model utilizes LSTM, attention mechanism, intermediate fusion, and label smoothing approaches to improve performance. The experiments were conducted on 474 video patients collected at Chulalongkorn University. The data set was divided into 134 depressions and 340 non-depressions. Our model achieves 91.67% accuracy, 91.40% precision, 87.03% recall, and 88.89% F1-score. In addition, our model is analyzed using an integrated gradient to explain the important facial features. The significant facial features related to depressive symptoms are head turning, no specific gaze, slow eye movement, no smiles, frowning, grumbling, and scowling, which express a lack of concentration, social disinterest, and negative feelings.

Field of Study: Computer Engineering

Student's Signature .....

Academic Year: 2023

Advisor's Signature .....

Co-advisor's Signature .....

## ACKNOWLEDGEMENTS

I extend my deepest gratitude to my academic advisor, Assoc. Prof. Dr. Peerapon Vateekul, for their unwavering support, guidance, and mentorship throughout the entire journey of this thesis. Their expertise and commitment to excellence have been instrumental in shaping not only the content but also the methodology of this research. I am truly fortunate to have had such a dedicated mentor.

I would also like to express my appreciation to my co-advisor, Assoc. Prof. Dr. Solaphat Hemrungronj, for their valuable insights and constructive feedback, which significantly enriched the quality of this work.

I want to express my appreciation to Mr. Kittipoch Saengsai and the team from the Center of Excellence in AI for Mental Health (AIMET), Chulalongkorn University, for providing the necessary resources and data collection. Their collaborative spirit and expertise greatly enhanced the overall quality of this research.

I would like to dedicate a special acknowledgment to my dear friends, Mr. Sothornin Mam, Mr. Passakron Phuangthongkham, Mr. Passin Pornvoraphat, Mr. Tanatorn Fajjaroenmongkol, Mr. Natch Sirisumpun, and Mr. Kittiwit Kumlungmak, whose unwavering support and friendship have been invaluable throughout the journey of this thesis.

This journey would have been far more challenging without the steadfast support of my friends and family. I owe a debt of gratitude for their understanding, encouragement, and love. Their unwavering belief in my abilities has been my source of strength.

Finally, I want to express my appreciation to the entire DataMind Lab at Chulalongkorn University community for providing the necessary resources, facilities, and academic infrastructure. The conducive atmosphere for learning and research has played a pivotal role in the successful completion of this thesis.

Yanisa Mahayossanunt

## TABLE OF CONTENTS

	Page
.....	iii
ABSTRACT (THAI).....	iii
.....	iv
ABSTRACT (ENGLISH).....	iv
ACKNOWLEDGEMENTS.....	v
TABLE OF CONTENTS.....	vi
LIST OF TABLES.....	ix
LIST OF FIGURES.....	x
CHAPTER 1 INTRODUCTION.....	1
1.1 Overview.....	1
1.2 Aims and Objectives.....	2
1.3 Scope of Work.....	2
1.4 Expected Benefits.....	2
1.5 Publication.....	3
CHAPTER 2 BACKGROUND KNOWLEDGE.....	4
2.1 Facial Action Coding System (FACS).....	4
2.2 OpenFace 2.2.0: Facial Behavior Analysis Toolkit.....	7
2.3 Machine Learning.....	10
2.3.1 Fusion Model.....	10
2.3.2 Long Short Term Memory (LSTM).....	12
2.3.3 Label Smoothing.....	13

2.3.4 Attention Mechanism .....	13
2.3.5 Transformer Model .....	15
2.3.6 Evaluation Measures.....	16
2.4 Integrated Gradient .....	18
CHAPTER 3 LITERATURE REVIEW .....	19
3.1 Facial Expressions and Depression Relation.....	19
3.2 Depression Detection Approaches.....	19
CHAPTER 4 METHODOLOGY .....	22
4.1 Facial Features Extraction.....	22
4.2 Input Preprocessing.....	22
4.2.1 Features Selection .....	22
4.2.2 Label Smoothing.....	25
4.3 Model Architecture .....	25
4.3.1 Baseline Fusion Bi-LSTM Model Architecture .....	25
4.3.2 Baseline Fusion Transformer Model Architecture .....	25
4.3.3 Individual Bi-LSTM Model Architecture .....	26
4.3.4 Early Fusion Bi-LSTM Model Architecture .....	27
4.3.5 Intermediate Fusion Bi-LSTM Model Architecture.....	28
4.3.6 Late Fusion Bi-LSTM Model Architecture .....	28
4.3.7 Individual Transformer Model Architecture .....	29
4.3.8 Early Fusion Transformer Model Architecture.....	29
4.3.9 Intermediate Fusion Transformer Model Architecture .....	30
4.3.10 Late Fusion Transformer Model Architecture .....	31
4.3.11 Individual Window Block LSTM Model Architecture .....	31



4.3.12	Early Fusion Window Block LSTM Model Architecture.....	32
4.3.13	Intermediate Fusion Window Block LSTM Model Architecture .....	33
4.3.14	Late Fusion Window Block LSTM Model Architecture.....	33
4.4	Integrated Gradient Explanation .....	34
CHAPTER 5 EXPERIMENTS AND RESULTS .....		35
5.1	Experimental Setups.....	35
5.1.1	Environment Detail .....	35
5.1.2	Data Distribution .....	35
5.1.3	Implementation.....	36
5.1.4	Evaluation.....	36
5.2	Experimental Results .....	37
CHAPTER 6 CONCLUSION AND FUTURE WORK.....		49
Conclusion.....		49
Future Work .....		49
REFERENCES .....		50
VITA.....		55

## LIST OF TABLES

	Page
Table 1 Action Units. ....	4
Table 2 Estimated Results of Openface.....	8
Table 3 Confusion Matrix.....	16
Table 4 Related Works of Depression Prediction.....	21
Table 5 Preliminary Experiment with Each Feature. Highlighted numbers refer to the winners.....	24
Table 6 Individual Bi-LSTM Model Hyperparameter. ....	27
Table 7 Individual Window Block LSTM Model Hyperparameter. ....	31
Table 8 Bi-LSTM Model Result. Highlighted numbers refer to the winners. ....	40
Table 9 Transformer Model Result. Highlighted numbers refer to the winners.....	41
Table 10 Window Block LSTM Model Result. Highlighted numbers refer to the winners.....	42
Table 11 Baseline Comparison. Highlighted numbers refer to the winners. ....	42
Table 12 Intermediate Fusion Bi-LSTM Model with Label Smoothing Result. Highlighted numbers refer to the winners.....	43
Table 13 Intermediate Fusion Window Block LSTM Model with Label Smoothing Result. Highlighted numbers refer to the winners.....	44
Table 14 Predicted Values of Intermediate Fusion Window Block LSTM Model with Label Smoothing (0.05, 0.95) .....	45

## LIST OF FIGURES

	<b>Page</b>
Figure 1 Facial Landmark Detection [8].....	7
Figure 2 Head Pose Tracking [8].....	7
Figure 3 Gaze Tracking [8].....	8
Figure 4 Facial Action Unit Recognition [8].....	8
Figure 5 Early Fusion.....	11
Figure 6 Intermediate/Joint Fusion.....	11
Figure 7 Late/Decision Fusion.....	12
Figure 8 Long Short Term Memory.....	12
Figure 9 Attention Model Diagram [20]. ....	14
Figure 10 Transformer Model Architecture [21].....	15
Figure 11 Facial Features Extraction Process. ....	22
Figure 12 Features Selection.....	23
Figure 13 Baseline Fusion Bi-LSTM Model Architecture. ....	25
Figure 14 Baseline Fusion Transformer Model Architecture.....	26
Figure 15 Individual Bi-LSTM Models Architectures.....	27
Figure 16 Early Fusion Bi-LSTM Model Architecture.....	27
Figure 17 Intermediate Fusion Bi-LSTM Model Architecture. ....	28
Figure 18 Late Fusion Bi-LSTM Model Architecture.....	28
Figure 19 Individual Transformer Model Architecture.....	29
Figure 20 Early Transformer Fusion Model Architecture. ....	30
Figure 21 Intermediate Transformer Fusion Model Architecture. ....	30

Figure 22 Late Transformer Fusion Model Architecture.....	31
Figure 23 Individual Window Block LSTM Model Architecture.....	32
Figure 24 Early Fusion Window Block LSTM Model Architecture.....	32
Figure 25 Intermediate Fusion Window Block LSTM Model Architecture.....	33
Figure 26 Late Fusion Window Block LSTM Model Architecture.....	33
Figure 27 Train Data Set, Dev Data Set, Test Data Set in 4 Class.....	35
Figure 28 Train Data Set, Dev Data Set, Test Data Set in 2 Class.....	36
Figure 29 Double Depression Class in Train Data Set.....	36
Figure 30 (A) Pose impact on model output, (B) Pose impact on model output magnitude, (C) Gaze impact on model output, (D) Gaze impact on model output magnitude, (E) AUr impact on model output, (F) AUr impact on model output magnitude, (G) AUc impact on model output, (H) AUc impact on model output magnitude. * Red color refers to a negative effect (tends to be non-depressive).....	46
Figure 31 (A) Positive/negative impact of all features, (B) Absolute impact (magnitude) of all features. * Red color refers to a negative effect (tends to be non-depressive).	47
Figure 32 Head Pose Movement [40].....	47
Figure 33 Gaze Movement.....	48

# CHAPTER 1

## INTRODUCTION

### 1.1 Overview

Major Depressive Disorder (MDD) is a common disorder in the global population. The impact of COVID-19 worsens the depression situation [1-5]. Patients who suffer from depression has rapidly increased during the pandemic. Depression is a fatal disorder that interferes with daily life and can lead to suicide. The prevention of life-threatening depression requires fast diagnosis and proper treatment. However, medical personnel are inadequate to measure depression for all citizens. One solution to reduce the responsibility of medical personnel that is currently utilized is the capability of artificial intelligence that can support medical personnel as a decision-support tool or primary decision tool.

Clinical interviews are one of the methods used to diagnose depression [6]. There is a list of questions to estimate mood, anhedonia (the inability to feel pleasure), anergia (a continual feeling of lack of energy), concentration, appetite, sleep, guilt, and suicide. To diagnose depression, a psychiatrist examines the patient's expression, posture, voice tone, and response content. Similarly, artificial intelligence has the capability of video, voice, and text processing that could potentially mimic a psychiatrist's observation.

Artificial intelligence has various approaches to processing data. Utilizing data from medical services necessitates obtaining patient consent, making it difficult to create large datasets. Therefore, a feature extraction tool is necessary to protect the privacy of patient information. Interview videos contain three types of data. There are expressions, voices, and textual content. The voice and textual content data are private because patients can be effortlessly identified by them. Hence, expression is a strategy to extract features for patient privacy protection.

The Facial Action Coding System (FACS) [7] defines a set of facial muscle movements that correspond to the displayed facial emotion. Facial expression features are extracted in this system to avoid identification. The OpenFace [8] tool takes responsibility for extracting features from interview videos. The extracted features are called Action Units (AUs). Therefore, the data set that was extracted from the tool is a time-series that contains a set of Action Units (AUs).

This thesis proposes deep learning approaches to time-series classification. In this research, we develop a fusion model to improve depression and non-depression classification from real-world data extraction and explain the result of the model in terms of facial key points.

## 1.2 Aims and Objectives

1. To provide methods that can improve the accuracy of the depression and non-depression classification models by utilizing time-series facial key point data extracted from interview videos to protect data privacy.
2. To explain the results of methods in terms of facial key point data.

## 1.3 Scope of Work

- Employ the dataset from the DMIND application, which is the result of a collaboration between Chulalongkorn University's faculties of medicine and engineering.
- Use facial key points that were extracted from the interview video as input.
- Develop neural network architectures for depression (moderate, severe) and non-depression (normal, mild) classifications.
- Evaluate the performance of the proposed neural network architectures in terms of classification.
- Explain the results of the proposed methods in terms of facial key points.

## 1.4 Expected Benefits

- Facial key point data can be used to differentiate between depression and non-depression.
- Facial expression video data can be made private by extracting time-series facial key point data.
- Patients can be helped to become aware of depression disorders.
- Medical personnel can be helped to reduce their workload.

- The insight of an explainable method can help people observe depression symptoms.

### 1.5 Publication

Mahayossanunt, Y.; Nupairoj, N.; Hemrungronj, S.; Vateekul, P. Explainable Depression Detection Based on Facial Expression Using LSTM on Attentional Intermediate Feature Fusion with Label Smoothing. *Sensors* **2023**, *23*, 9402. <https://doi.org/10.3390/s23239402>



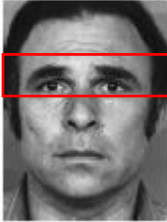



## CHAPTER 2

### BACKGROUND KNOWLEDGE




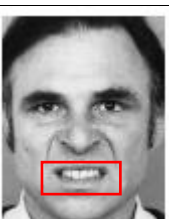



#### 2.1 Facial Action Coding System (FACS)





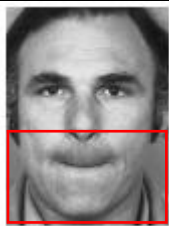

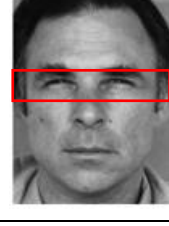
The Facial Action Coding System (FACS) [7] is a system that describes facial muscle movements as action units. Table 1 lists all of the action units employed in this thesis. This system includes gaze direction and head pose. The original creator is Carl-Herman Hjortsjö who created 23 facial motion units in 1970. Paul Ekman, and Wallace Friesen continued to develop this system after it was first published in 1978 and substantially updated in 2002.

Table 1 Action Units.

Action Unit	Description	Example
1	Inner Brow Raise	
2	Outer Brow Raise	
4	Brow Lowerer	
5	Upper Lid Raise	



6	Cheek Raise	
7	Lids Tight	
9	Nose Wrinkle	
10	Upper Lip Raiser	
12	Lip Corner Puller	
14	Dimpler	
15	Lip Corner Depressor	

17	Chin Raiser	
20	Lip Stretch	
23	Lip Tightener	
25	Lips Part	
26	Jaw Drop	
28	Lip Suck	
45	Blink	

## 2.2 OpenFace 2.2.0: Facial Behavior Analysis Toolkit

OpenFace [8] is an open source framework that provides facial land mark detection [9] in Figure 1, head pose tracking [10] in Figure 2, eye gaze [11] in Figure 3 and facial action unit estimation [12] in Figure 4. Table 2 shows the results of the Openface tracking estimation. As a result, the tracking values provided by Openface cannot achieve 100% accuracy. Therefore, the maximum tracker's confidence value is 98%.



จุฬาลงกรณ์มหาวิทยาลัย  
Figure 1 Facial Landmark Detection [8].  
CHULALONGKORN UNIVERSITY

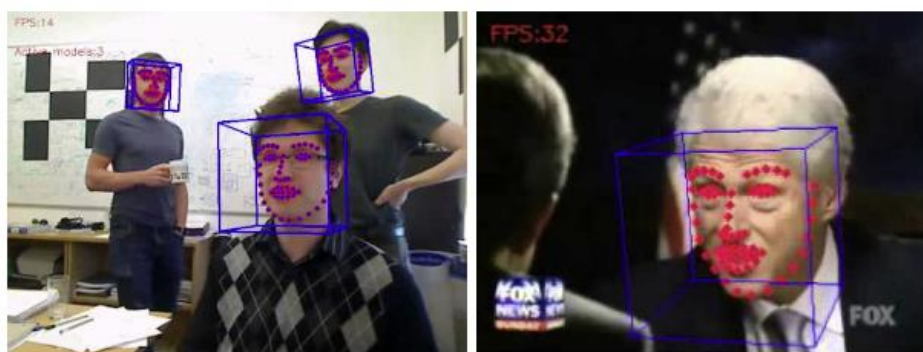


Figure 2 Head Pose Tracking [8].

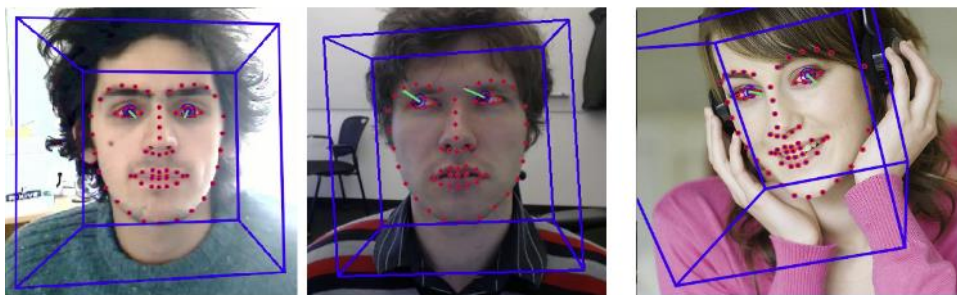


Figure 3 Gaze Tracking [8].

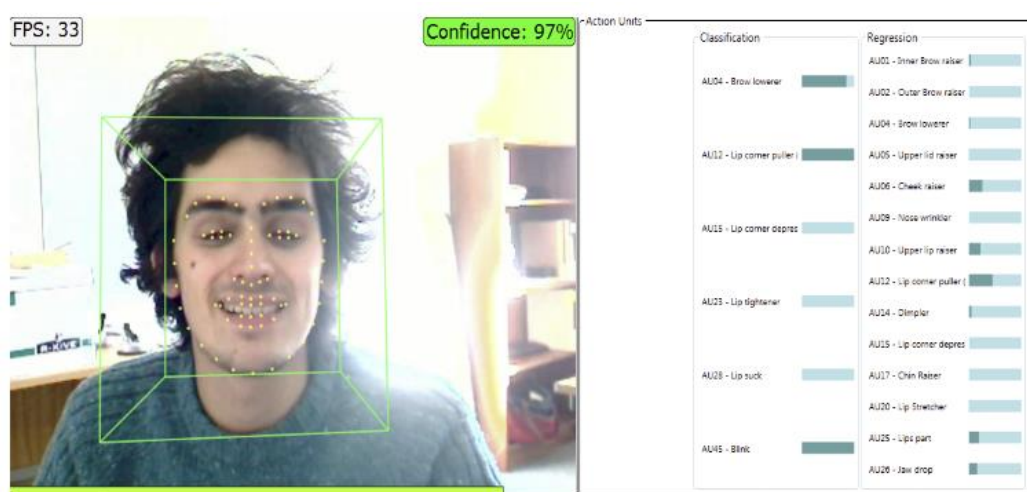


Figure 4 Facial Action Unit Recognition [8].

Table 2 Estimated Results of Openface.

\* CCC refers to concordance correlation coefficient.

Openface	Mean Absolute Error
Head pose estimation results [8] on the BU dataset [13]	2.6
Head pose estimation results [8] on ICT-3DHP dataset [14]	3.2
Gaze estimation results [8] on MPIIGaze dataset [15]	9.1
Action units estimation results [8] on DISFA validation set [16]	CCC 0.73

The output format of the Openface tool is an extracted features file that contains time-series basic information, head pose tracking, gaze tracking, and facial action units. The following is a description of header information.

### Basic Information Section

- frame: the number of the frame (in the case of sequences).
- face\_id: the face id (in case of multiple faces)
- timestamp: the timer of video being processed in seconds (in case of sequences)
- confidence: the tracker's confidence in its current landmark detection estimate.
- success: the track is successful.

### Pose Tracking Section

- pose\_Tx: the horizontal location of the head with respect to the camera in millimeters.
- pose\_Ty: the vertical location of the head with respect to the camera in millimeters.
- pose\_Tz: the millimeter distance between the head and the camera.
- pose\_Rx: rotation is in radians around the X axis (pitch), a left-handed positive sign.
- pose\_Ry: rotation is in radians around the Y axis (yaw), a left-handed positive sign.
- pose\_Rz: rotation is in radians around the Z axis (roll), a left-handed positive sign.

### Gaze Tracking Section

- gaze\_0\_x: x eye gaze direction vector in world coordinates for the leftmost eye
- gaze\_0\_y: y eye gaze direction vector in world coordinates for the leftmost eye
- gaze\_0\_z: z eye gaze direction vector in world coordinates for the leftmost eye
- gaze\_1\_x: x eye gaze direction vector in world coordinates for the rightmost eye
- gaze\_1\_y: y eye gaze direction vector in world coordinates for the rightmost eye
- gaze\_1\_z: z eye gaze direction vector in world coordinates for the rightmost eye
- gaze\_angle\_x: eye gaze direction in radians in world coordinates from left to right (from positive to negative)
- gaze\_angle\_y: eye gaze direction in radians in world coordinates from up to down (from positive to negative)

## Facial Action Units

The system can detect the intensity (from 0 to 5) of 17 AUs:

AU01\_r, AU02\_r, AU04\_r, AU05\_r, AU06\_r, AU07\_r, AU09\_r, AU10\_r, AU12\_r, AU14\_r, AU15\_r, AU17\_r, AU20\_r, AU23\_r, AU25\_r, AU26\_r, AU45\_r

And the presence (0 absent and 1 present) of 18 AUs:

AU01\_c, AU02\_c, AU04\_c, AU05\_c, AU06\_c, AU07\_c, AU09\_c, AU10\_c, AU12\_c, AU14\_c, AU15\_c, AU17\_c, AU20\_c, AU23\_c, AU25\_c, AU26\_c, AU28\_c, AU45\_c

## 2.3 Machine Learning

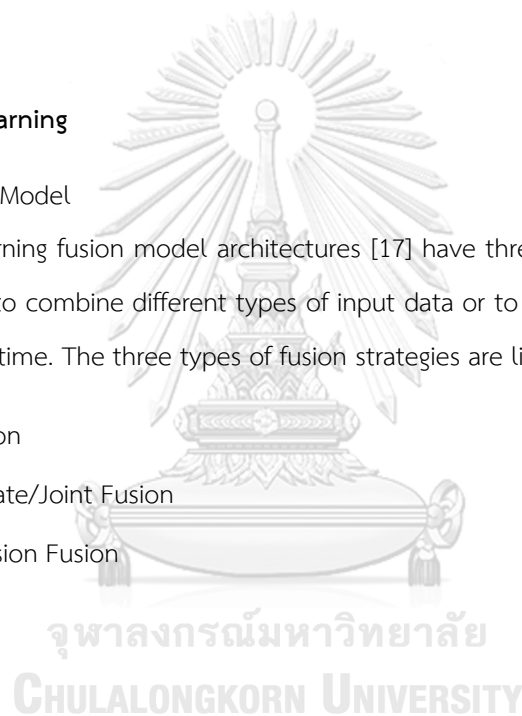
### 2.3.1 Fusion Model

Machine learning fusion model architectures [17] have three types of strategies. A fusion model can be used to combine different types of input data or to run multiple machine learning models at the same time. The three types of fusion strategies are listed below.

- Early Fusion
- Intermediate/Joint Fusion
- Late/Decision Fusion

#### Early Fusion

The goal of early fusion is to combine data before putting it into a model. Combined data can be original data or features extracted from raw data. There are various combinatorial methods. In a neural network, data combining typically occurs through a concatenation layer or pooling layer. Figure 5 depicts an early fusion model architecture.



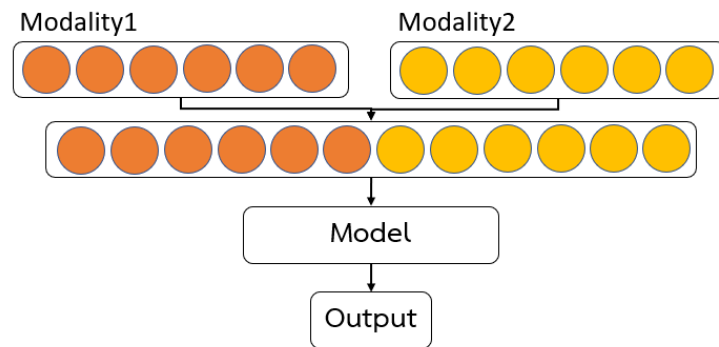


Figure 5 Early Fusion.

### Intermediate/Joint Fusion

The combination of output from multiple neural networks before making a decision is known as intermediate/joint fusion. This strategy can update weights for all neural networks because the loss from the model can be propagated back to multiple neural networks. Figure 6 shows an example of an intermediate/joint fusion model's architecture.

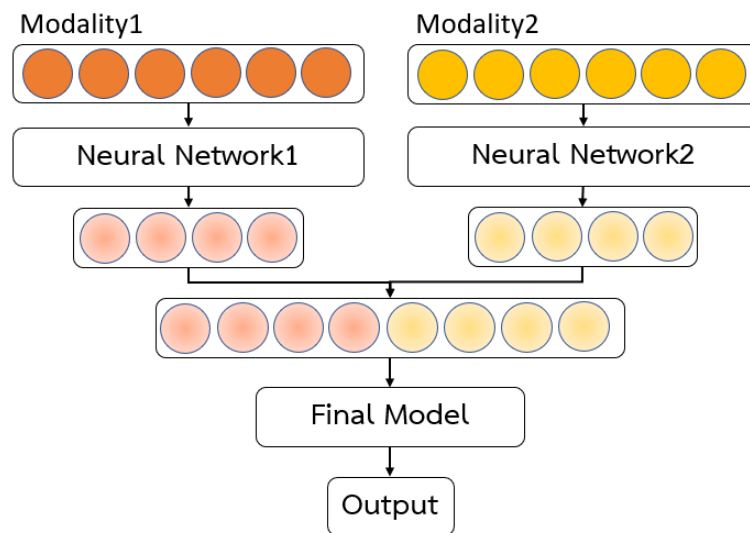


Figure 6 Intermediate/Joint Fusion.



## Late/Decision Fusion

Late/Decision Fusion is the predictions of multiple model aggregation at the decision level. This fusion strategy can be called decision fusion. There are various aggregation techniques, for instance, majority voting, averaging, and weight voting.

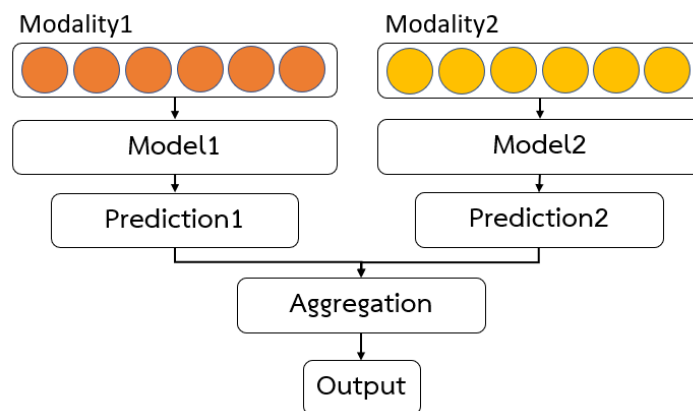


Figure 7 Late/Decision Fusion.

### 2.3.2 Long Short Term Memory (LSTM)

Long Short Term Memory (LSTM) [18] a type of recurrent neural network that can partially solve the vanishing gradient problem in recurrent neural networks. LSTM has a cell state and gate to control data flow. Figure 8 shows a long short term memory diagram.

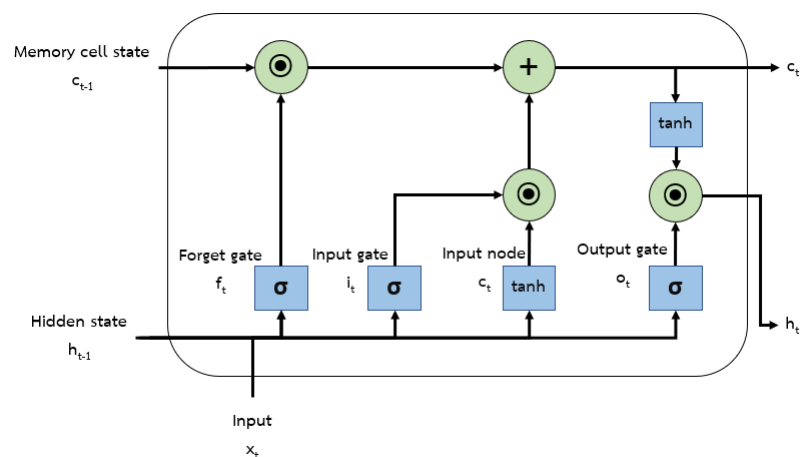


Figure 8 Long Short Term Memory.



### 2.3.3 Label Smoothing

Label smoothing is a technique in machine learning that prevents models from becoming overconfident. This technique can increase robustness and improve the classification model. The following is the definition of a soft label and a hard label. [19].

- A soft label is a score that has some probability or likelihood attached to it. For instance, [0.2, 0.8].
- A hard label is typically classified into one of two categories. It is binary in nature (either 0 or 1).

The formula for label smoothing that transforms a hard label into a soft label is shown in ( 1 )

$$y_k^{LS} = y_k(1-\alpha) + \alpha/K \quad (1)$$

Where:

- $y_k^{LS}$  is a soft label.
- $y_k$  is a hard label.
- $\alpha$  is a label smoothing that should be in range 0 to 1.
- $K$  is a number of classes.

### 2.3.4 Attention Mechanism

Attention mechanism [20] is a technique that focuses the model's attention on a specific point. This method can improve the model's performance and be used for explaining predicted results. Figure 9 shows an attention model diagram. The bidirectional LSTM generates a sequence of forward and backward hidden states in the encoder ( 2 ). The context vector is calculated by weighting the hidden states ( 3 ). Each hidden state is weighted by  $\alpha_{tj}$ . The weight  $\alpha_{tj}$  (the alignment score) is computed by a softmax function ( 4 ). The score function used in the alignment score uses tanh as a non-linear activation function,  $v_\alpha$  and  $W_\alpha$  as the weight matrices ( 5 ).

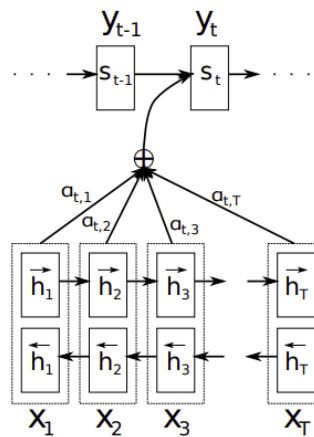


Figure 9 Attention Model Diagram [20].

Hidden States

$$h_i = [\vec{h}_i; \overleftarrow{h}_i]$$

(2)

Context Vector

$$c_i = \sum_{i=1}^{T_x} \alpha_{ti} h_i$$

(3)

Alignment Score

$$a_{t,i} = \text{align}(y_t, x_i) = \frac{\exp(\text{score}(s_{t-1}, h_i))}{\sum_{i=1}^n \exp(\text{score}(s_{t-1}, h_i))}$$

(4)

Score Function

$$\text{score}(s_t, h_i) = v_\alpha^T \tanh(W_\alpha [s_t; h_i])$$

(5)

### 2.3.5 Transformer Model

Transformer Model [21] is a novel architecture that is used to solve sequence-to-sequence tasks with long-range dependencies. The transformer model utilizes attention to handle the dependencies between input and output. Figure 10 shows the transformer model architecture, in which the encoder block has one layer of multi-head attention followed by a feed-forward neural network.

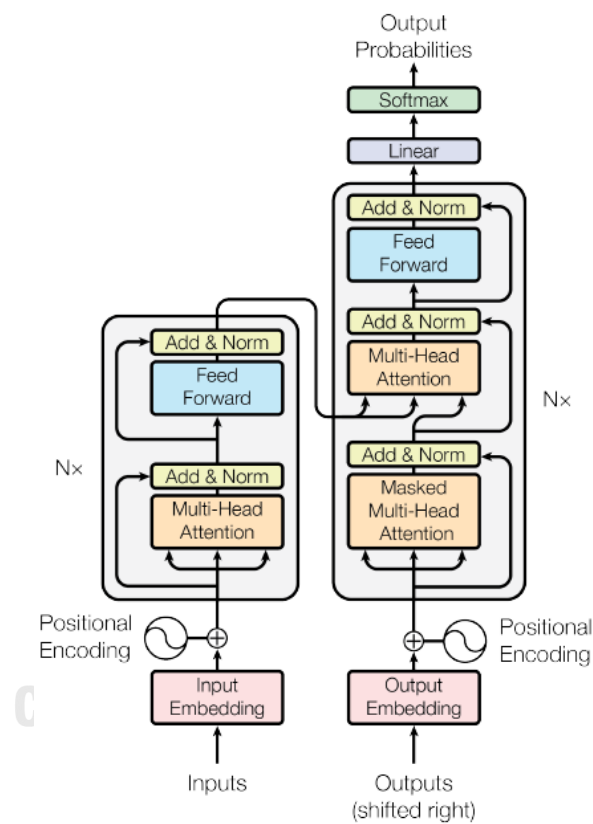


Figure 10 Transformer Model Architecture [21].

### 2.3.6 Evaluation Measures

The following are some common performance metrics for classification:

- Accuracy
- Confusion Matrix
- Precision
- Recall
- F1 score

#### Accuracy

Accuracy is the ratio between the number of correctly predicted results and the total number of results. Accuracy formular shows in ( 6 ).

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (6)$$

#### Confusion Matrix

Confusion matrix is a table that contains different combinations of predicted and actual values. The values are true positive value, true negative value, false positive value, and false negative value. Table 3 shows the confusion matrix.

Table 3 Confusion Matrix.

	Positive Actual Values	Negative Actual Values
Positive Predicted Values	True Positive Value (TP)	False Positive Value (FP)
Negative Predicted Values	False Positive Value (FP)	False Negative Value (FN)

### Precision

Precision is a measure of relevant instances among the retrieved instances. It can be called positive predictive value. Precision formular is shown in ( 7 ).

$$\text{Precision} = \frac{\text{True Positive Value}}{\text{True Positive Value} + \text{False Positive Value}} \quad (7)$$

### Recall

Recall is a measure of the relevant instances that were retrieved. It can be called sensitivity. Recall formular is shown in ( 8 ).

$$\text{Recall} = \frac{\text{True Positive Value}}{\text{True Positive Value} + \text{False Negative Value}} \quad (8)$$

### F1 Score

F1 score is a measure of model accuracy on a dataset. It is defined as the harmonic mean of precision and recall. F1 score formular shows in ( 9 ).

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

## 2.4 Integrated Gradient

Integrated Gradient (IG) [22] is an interpretability technique used in machine learning and deep learning models to visualize the input features and model predictions. The advantage of this technique is that the original deep neural networks are not modified while applying IG.

The integrated gradient technique is to compute the integral of the gradients of the model's predictions with respect to the input features along the straight-line path from a baseline (zero input) to the input being interpreted. By integrating the gradients along this path, Integrated gradients assigns an important score to each feature, indicating how much it contributes to the model's prediction for a specific input. The formula for computing the integrated gradient for a particular input feature  $i$  shows in ( 10 )

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(z(\alpha))}{\partial x_i} d\alpha \quad (10)$$

Where:

- $IG_i(x)$  is the integrated gradient for the  $i$  feature of the input  $x$
- $x_i$  is the value of the  $i$  feature in the input being interpreted.
- $x'_i$  is the value of the  $i$  feature in the baseline input.
- $F(z(\alpha))$  is the model's prediction function, where  $z(\alpha)$  is the interpolated input along the path from the baseline to the actual input, defined as
 
$$z(\alpha) = x' + \alpha(x - x')$$
- The integral term represents the partial derivative of the model's prediction with respect to the  $i$  feature, evaluated at the interpolated inputs  $z(\alpha)$ .

## CHAPTER 3

### LITERATURE REVIEW

#### 3.1 Facial Expressions and Depression Relation

Facial expressions can show human emotion that is associated with depression symptoms. During a depression diagnosis interview, the patients' facial expressions represent their feelings and emotions, which can be used in depression classification. In the medical field, there are studies about facial expressions, depression, and brain network relationships. In [23], they experiment with the effects of positive and negative facial expressions on electroencephalographic (EEG) analysis. The results show that facial expression can be used to identify the side of the facial muscles in EEG analysis. In [24], they experiment with the effect of happy and sad facial expression reactions in depressed patients and non-depressed volunteers by using functional magnetic resonance imaging (fMRI). The results show that depressed patients respond to sad facial expressions more than normal people and respond to happy facial expressions less than normal people. In [25], the results confirmed that neural activity in the cerebellum from fMRI scans has a relationship with depression. The study, as previously described, confirms that facial expression is related to the brain network via EEG and MRI observations. Furthermore, [26] shows that facial modality is associated with voice modality in emotion expression, and the experimentation in [27] shows that humans can distinguish depression symptoms from facial expressions. In the same direction, [28] proves that depression can be predicted by using face and eye movement tracking during a cognitive task. As a result, depression symptoms manifest as intensities of reduced mouth or eye movements at various stages of a cognitive task. Therefore, the evidence that facial expressions are related to depression symptoms exists today.

#### 3.2 Depression Detection Approaches

Artificial intelligence technology rapidly enhances various fields. The medical field is the one that exploits this technology to improve medical performance. Diagnosis is a popular area in which AI can play a role as a pre-diagnosis or decision-support tool because it improves the speed and accuracy of the diagnosis process. Currently, there are several techniques in artificial intelligence to detect psychiatric disorders [29]. There are three main categories of raw data that

become the input of the detected model: MRI, EGG, and kinesics diagnosis (including behavioral, facial, and other physical data). Algorithms that are used with this data can be categorized into five types. There are Bayesian models, logistic regression, decision trees, support vector machines, and deep learning. In this thesis, we focus on depression detection using facial expressions. Therefore, facial cues that express depression symptoms are pupil dilation, action units, facial expressions (emotion), and head pose [30].

Researchers have recently focused on the facial modality. They experiment with facial modality alone or in combination with other modality to predict depression. In [31], multi-model fusion of visual, voice, and text is proposed with a concordance correlation coefficient (CCC) of 0.67 in the E-DAIC dataset. In [32] they propose a method to reduce AUs in a feed-forward neural network (FFNN) by using Particle Swarm Optimization (PSO) to select the best predictors of AUs. The best predictors are AU04\_r, AU06\_r, AU09\_r, AU10\_r, AU15\_r, AU25\_r, AU26\_r, AU04\_c, AU12\_c, AU23\_c, AU28\_c, and AU45\_c in the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC WOZ) data set with 97.83% accuracy. In [33] propose a facial and voice fusion transformer network to estimate depressive levels. They categorize the depression score from PHQ-8 into five levels for use as the first classification label in multi-task learning. The second multi-task learning label is the PHQ-8 regression label. Their proposed method achieves a CCC of 0.733 in the E-DAIC data set. In [34], they propose Fisher Discriminant Ratio (FDR) and Incremental Linear Discriminant Analysis (ILDA) to reduce and select facial features from the DAIC WOZ dataset. Their method achieves an F1 score of 80.5%, the highest score in the DAIC WOZ dataset. In [35], they utilize deep learning to classify posttraumatic stress disorder (PTSD) and major depressive disorder (MDD) based on facial features, movement intensity, speech, and content. This raw data was collected from 81 patients in one month. The results show that the PTSD classification reached 90% accuracy and the MDD classification reached 86% accuracy. All studies aim to improve depression classification performance. The conclusion of this study is shown in Table 4.

According to Table 4, depression prediction using facial features in the artificial intelligence field appears to be gaining popularity in recent years. Therefore, there are various possibilities to explore for improving the performance of the model.



Table 4 Related Works of Depression Prediction.

\* CCC refers to concordance correlation coefficient.

Year	Techniques	Data sets	Questionnaires	Accuracy	Precision	Recall	F1 score
2019 [31]	Multi-Model, Bi-LSTM	E-DAIC	PHQ-8	CCC 0.67	-	-	-
2021 [32]	PSO, FFNN	DAIC WOZ	PHQ-8	97.83%	-	-	-
2021 [33]	Multi-Modal Transformer	E-DAIC	PHQ-8	CCC 0.733	-	-	-
2022 [34]	FDR, ILDA	DAIC WOZ	PHQ-8	-	-	-	80.5%
2022 [35]	FFNN	Their own	-	86%	83%	82%	82%

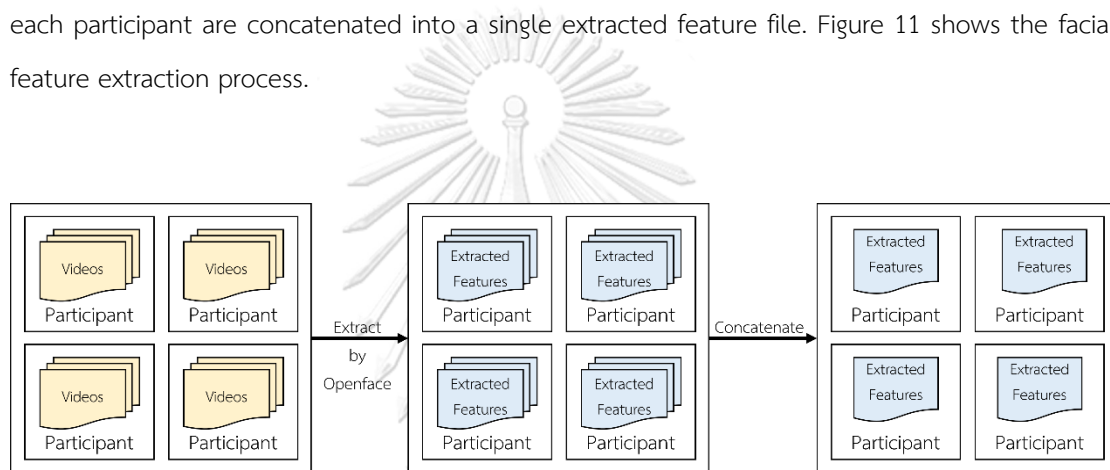
In all related works, the experiment models in [31, 33] can be compared with ours because the E-DAIC data set is extracted from OpenFace version 2, the same as ours. On the other hand, the experiment models in [32, 34] use the DAIC WOZ data set that extracted facial features from OpenFace version 1, which is different from ours. The model in [31] will be the baseline of our Bi-LSTM model. The model in [33] will be the baseline of our transformer model.

## CHAPTER 4

### METHODOLOGY

#### 4.1 Facial Features Extraction

Raw data from DMIND applications contains different numbers of videos depending on how participants answer the list of DMIND interview questions. The OpenFace tool extracts head pose, gaze, and action units from all participant videos. The extracted data is a time-series extracted features file that contains head pose, gaze, and action units per video. However, input should be one extracted feature file per participant. As a result, the extracted feature files from each participant are concatenated into a single extracted feature file. Figure 11 shows the facial feature extraction process.



*Figure 11 Facial Features Extraction Process.*

#### 4.2 Input Preprocessing

##### 4.2.1 Features Selection

Extracted features files contain head pose, gaze, and action units. Each facial feature (head pose, gaze, and action units) has a sub-feature group as follows: the head pose feature has location and rotation sub-feature groups. The gaze feature has vector and radian sub-feature groups. Sub-feature groups for action units include intensity and presence. We separate data from extracted feature files by group and standardize all groups by removing the mean and scaling to unit variance (Standard Scaler). As a result, for input into models, facial features are divided into six sub-feature groups. The head rotation sub-feature groups and head location sub-feature groups are repeated with different units, the same as the gaze vector sub-feature groups and gaze radian sub-feature groups, as shown in Figure 12. We experimented with them using a single model to compare their results. The results, as shown in Table 5, indicated that the head pose

location and gaze vector direction have poor performance in classifying depression. Therefore, the head pose location and gaze vector direction sub-feature groups are not selected for experimentation to eliminate redundant data and model size. The action units presence sub-feature groups and the action units intensity sub-feature groups, all of which use distinct estimation models. Both sub-feature groups are selected. The features selection is shown in Figure 12. The list of sub-feature groups is summarized as follows:

1. Pose\_R Head pose rotation has 3 features.
2. Gaze\_Angle: Gaze angle direction has 2 features.
3. AU\_r: Action unit intensity has 17 features.
4. AU\_c: Action unit presence has 18 features.

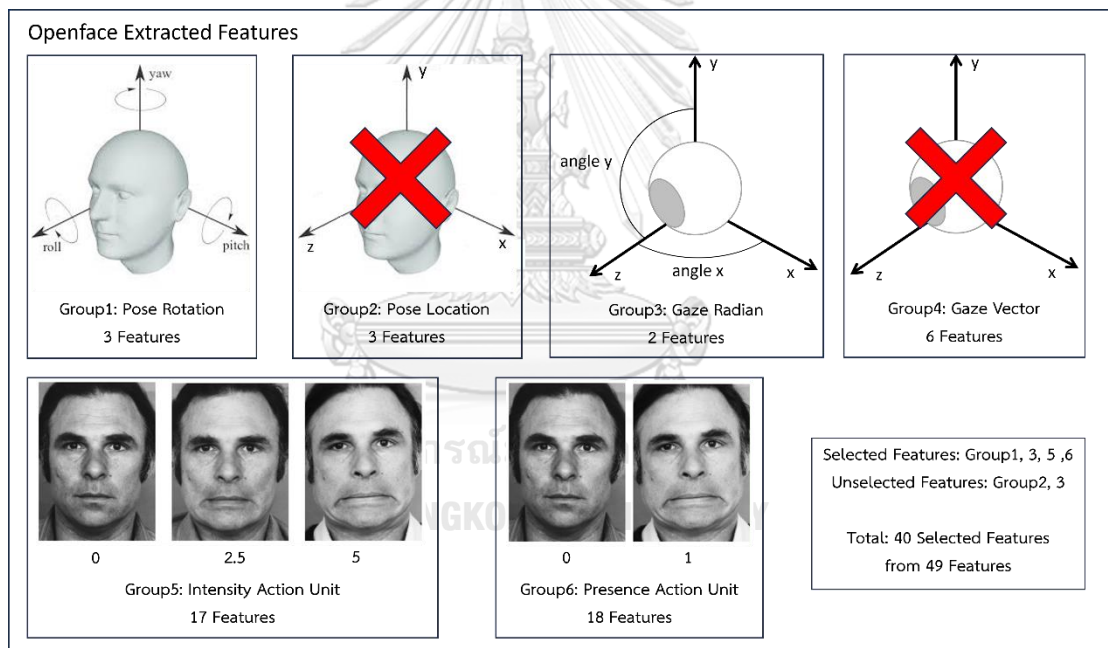


Figure 12 Features Selection.

The longest time-series of inputs is not over 11 minutes. Therefore, this number is selected to be calculated as timesteps. Timestep is calculated from time (minutes) \* 60 (seconds) \* 30 (frames/second). In this thesis, the timestep is 19800. To fit the timestep, inputs that do not reach the timestep are padded with zero.

Table 5 Preliminary Experiment with Each Feature. Highlighted numbers refer to the winners.

Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
Pose_R	6	7	32	3	79.17	74.36	68.79	70.52	77.88	79.17	77.84
Pose_L	5	8	33	2	79.17	75.96	66.37	68.42	78.03	79.17	76.86
Gaze_R	8	5	31	4	81.25	76.39	75.05	75.66	80.84	81.25	81.01
Gaze_V	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
AU_r	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27
AU_c	8	5	33	2	85.42	83.42	77.91	79.99	84.99	85.42	84.77



#### 4.2.2 Label Smoothing

Label smoothing is applied to improve the model's performance because it can prevent a model from becoming overconfident in its predictions. The model can be improved by label smoothing because depressive syndrome is not clearly defined, especially at a mild and moderate level. We apply a label smoothing number with a range of 0 to 0.9 with 0.1 increments solely on the best model to reduce computation time.

### 4.3 Model Architecture

#### 4.3.1 Baseline Fusion Bi-LSTM Model Architecture

The baseline fusion Bi-LSTM model [31] passes pose, gaze, and facial action units through its own single layer of 200 Bi-LSTM cells. Their output is concatenated before passing through the attention layer. The output of attention is passed through another Bi-LSTM with 200 cells, followed by global max pooling. After global max pooling, the output is passed through a feed-forward network with 128 hidden units. The total number of parameters is 888,786. This model is shown in Figure 13.

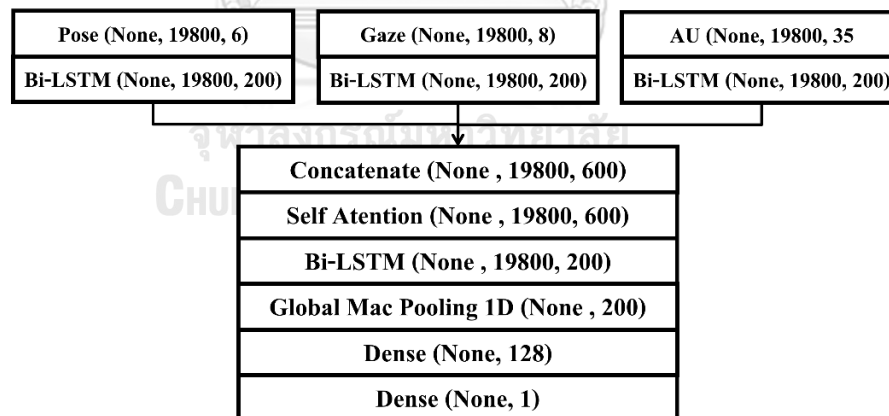


Figure 13 Baseline Fusion Bi-LSTM Model Architecture.

#### 4.3.2 Baseline Fusion Transformer Model Architecture

The baseline fusion transformer model [33] makes use of the early fusion technique. All features (pose, gaze, and action units) concatenate before passing through the model. Unfortunately, we cannot use 2048 timesteps as efficiently as [33] because of the environment.

The timesteps that are used in this baseline are reduced to 1320 timesteps (average of 30 frames per second to 2 frames per second). The multi-head attention number is set to 1, the feed-forward layer's hidden dimension is set to 2048, and the number of transformer encoders is set to 6. After that, a rectified linear unit (ReLU) is applied. However, units that used this layer are not mentioned. Therefore, we use 32 units for ReLU. The total number of parameters is 1,831,270. This model is shown in Figure 14.

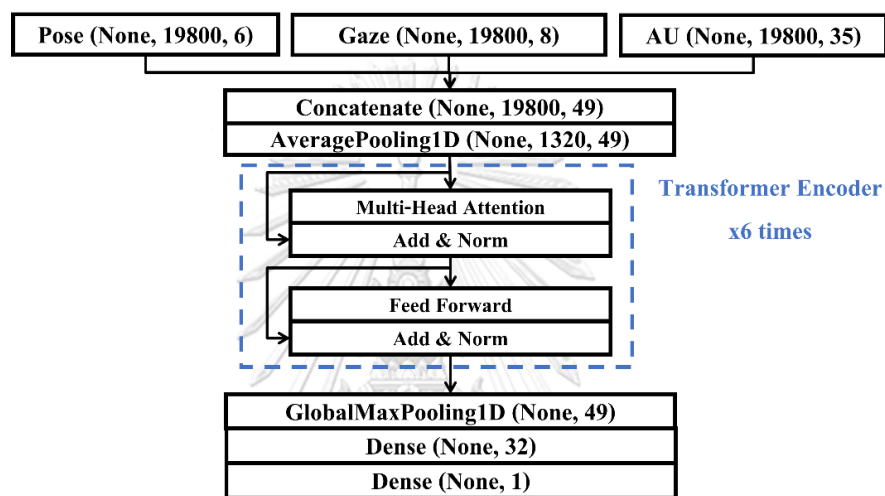


Figure 14 Baseline Fusion Transformer Model Architecture.

#### 4.3.3 Individual Bi-LSTM Model Architecture

Bidirectional LSTM and self-attention are used to generate individual Bi-LSTM model architectures for four sub-feature groups. Four models are produced with the same layers that are shown in Figure 15. Hyperparameters are set for Pose\_R, Gaze\_Angle, AU\_r ,and AU\_c respectively, as described in Table 6. The hidden units, units1, and units2, are powers of two. The selected units are the best values from hyperparameter tuning techniques.

Table 6 Individual Bi-LSTM Model Hyperparameter.

Hyperparameter	Pose_R	Gaze_Angle	AU_r	AU_c
Features	3	2	17	18
Hidden_units	64	16	128	128
Units1	32	8	64	64
Units2	16	4	32	32
Total parameters	31,722	20,702	72,314	72,826

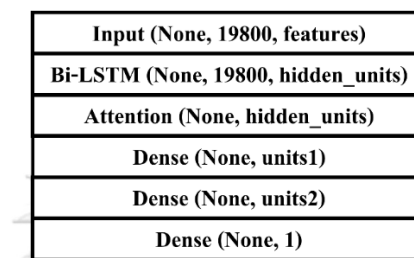


Figure 15 Individual Bi-LSTM Models Architectures.

#### 4.3.4 Early Fusion Bi-LSTM Model Architecture

Early Fusion Bi-LSTM Model Architecture concatenates four sub-feature groups into one group. As a result, a model receives one input that contains four sub-features with 40 features. The model is similar to an individual model. The total number of parameters is 84,090. This architecture is shown in Figure 16. All hidden units are powers of two.

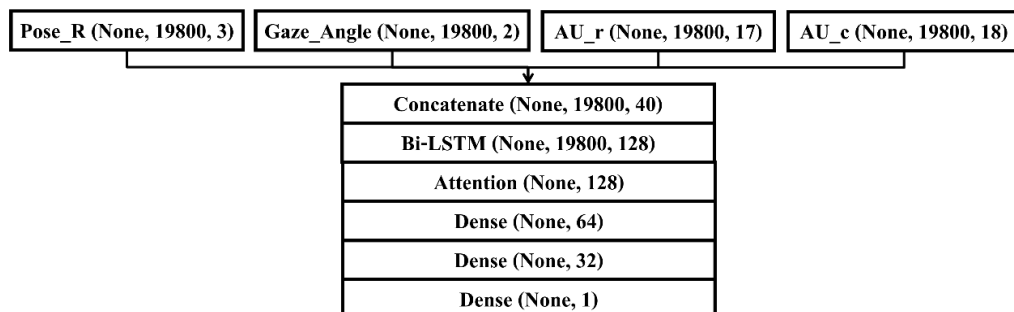


Figure 16 Early Fusion Bi-LSTM Model Architecture.

#### 4.3.5 Intermediate Fusion Bi-LSTM Model Architecture

Intermediate Fusion Bi-LSTM Architecture utilizes four Bi-LSTM individual models by removing the aggregation section. Output from the attention layer of four individual models is concatenated before the decision layer. The total number of parameters is 225,458. This architecture is shown in Figure 17. All hidden units are powers of two. The selected units are the best values from hyperparameter tuning techniques.

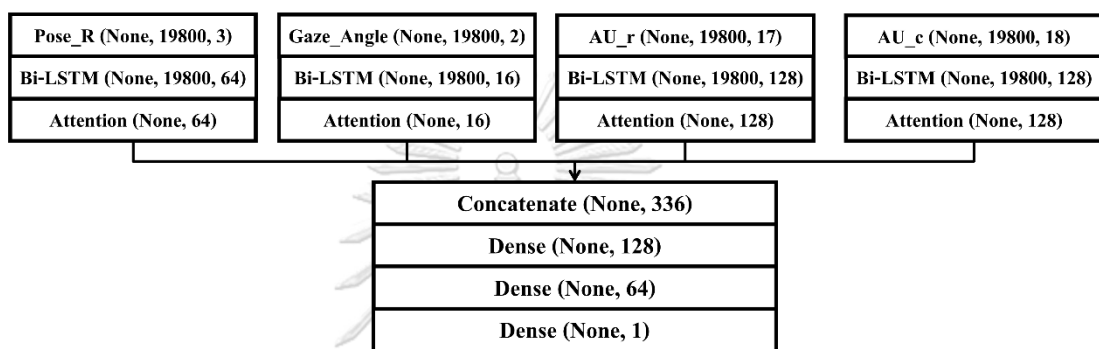


Figure 17 Intermediate Fusion Bi-LSTM Model Architecture.

#### 4.3.6 Late Fusion Bi-LSTM Model Architecture

Late Fusion Bi-LSTM Model Architecture utilizes four individual models to determine depression and average their results. The total number of parameters is 226,556. Figure 18 depicts the method of aggregation.

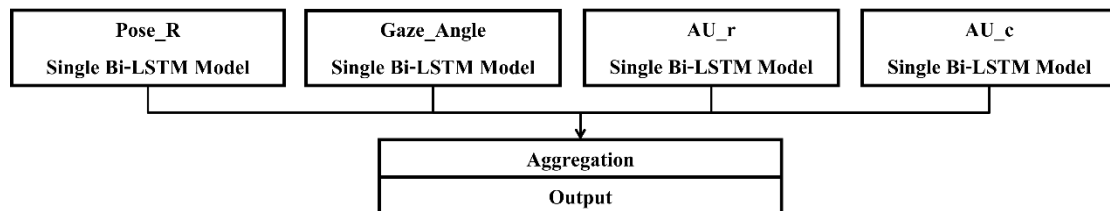


Figure 18 Late Fusion Bi-LSTM Model Architecture.



#### 4.3.7 Individual Transformer Model Architecture

Individual Transformer Models Architectures for four sub-feature groups utilize a transformer encoder, followed by a global average pooling layer and a dense layer. Before the multi-head attention layer, average pooling that has pool size 15 and stride 15 (average 30 frames per second to 2 frames per second) is applied to reduce timesteps because of memory. Four models are produced with the same layers that are shown in Figure 19 except for features. Features are 3, 2, 17, and 18 for Pose\_R, Gaze\_Angel, AU\_r, and AU\_c, respectively. The total number of parameters for Pose\_R, Gaze\_Angel, AU\_r, and AU\_c are 22,042, 15,890, 108,170, and 114,322, respectively.

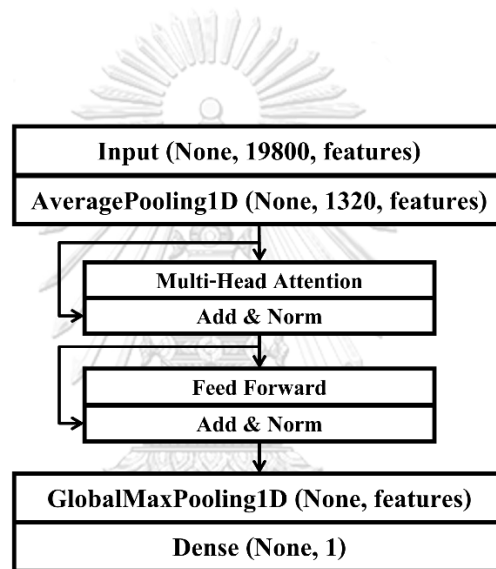


Figure 19 Individual Transformer Model Architecture.

#### 4.3.8 Early Fusion Transformer Model Architecture

Early Fusion Transformer Model Architecture concatenates four sub-feature groups into one group. As a result, a model receives one input that contains four sub-features with 40 features. The model is similar to an individual model. The total number of parameters is 249,666. This architecture is shown in Figure 20.

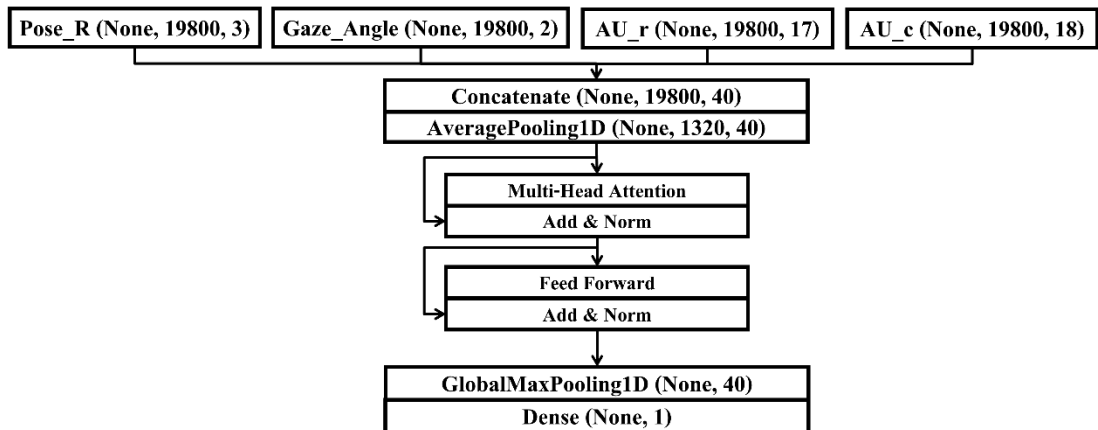


Figure 20 Early Transformer Fusion Model Architecture.

#### 4.3.9 Intermediate Fusion Transformer Model Architecture

Intermediate Fusion Transformer Model Architecture utilizes four individual transformer models by removing the aggregation section. Output from the global average pooling layer of four individual models is concatenated before the decision layers. The total number of parameters is 260,418. This architecture is shown in Figure 21.

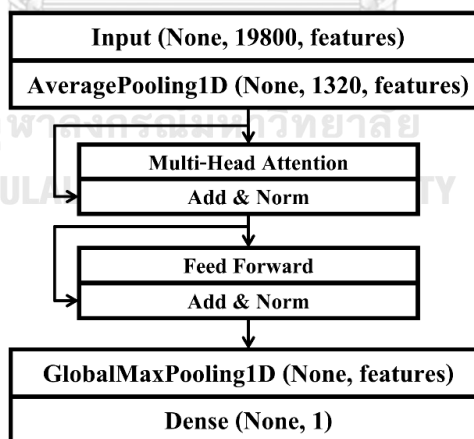


Figure 21 Intermediate Transformer Fusion Model Architecture.

#### 4.3.10 Late Fusion Transformer Model Architecture

Late Fusion Transformer Model Architecture employs four separate transformer models to determine depression and average their decision to results. The total number of parameters is 260,424. The aggregation method is shown in Figure 22.

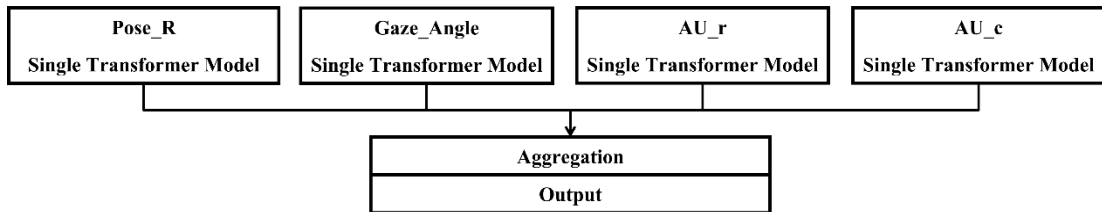


Figure 22 Late Transformer Fusion Model Architecture.

#### 4.3.11 Individual Window Block LSTM Model Architecture

Individual Window Block LSTM Model Architecture is utilized: reshape layer, time distribution with LSTM layer, time distribution with attention, attention layer, and feed forward layers. The reshape layer is utilized for converting 19800 frames to 30 frames x 660 second. The model hyperparameter and architecture are shown in Table 7 and Figure 23, respectively. The hidden units, units1, and units2, are powers of two. The selected units are the best values from hyperparameter tuning techniques.

จุฬาลงกรณ์มหาวิทยาลัย  
CHULALONGKORN UNIVERSITY

Table 7 Individual Window Block LSTM Model Hyperparameter.

Hyperparameter	Pose_R	Gaze_Angle	AU_r	AU_c
Features	3	2	17	18
Hidden_units	32	32	64	64
Units1	16	16	32	32
Units2	8	8	16	16
Total parameters	6,035	5,907	24,435	24,691

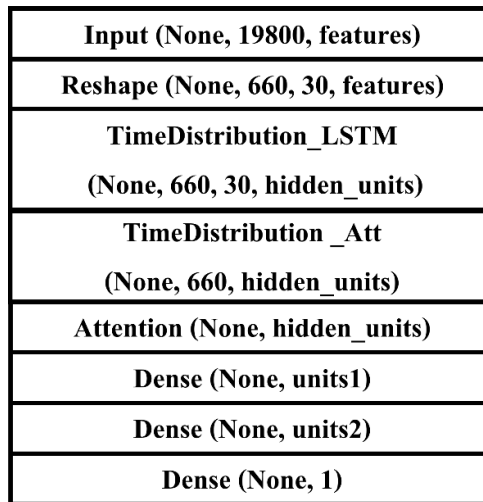


Figure 23 Individual Window Block LSTM Model Architecture.

#### 4.3.12 Early Fusion Window Block LSTM Model Architecture

Early Fusion Window Block LSTM Model Architecture utilizes a concatenate layer to concatenate all input features before passing through the following layers. The following layers are: reshape layer, time distribution with LSTM layer, time distribution with attention layer, attention layer, and feed forward layers. The total number of parameters is 30,323. The model architecture is shown in Figure 24. All hidden units are powers of two.

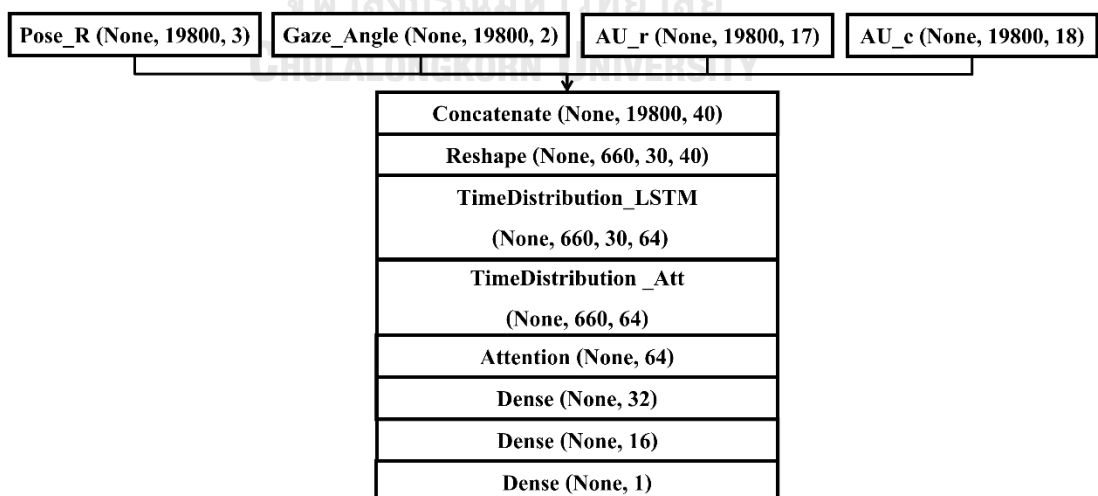


Figure 24 Early Fusion Window Block LSTM Model Architecture.

#### 4.3.13 Intermediate Fusion Window Block LSTM Model Architecture

Intermediate Fusion Window Block LSTM Model Architecture utilizes a concatenate layer before feed-forward layers. Each input feature passes through a reshape layer, a time distribution with an LSTM layer, a time distribution with an attention layer, and an attention layer. The total number of parameters is 90,057. The model architecture is shown in Figure 25. All hidden units are powers of two. The selected units are the best values from hyperparameter tuning techniques.

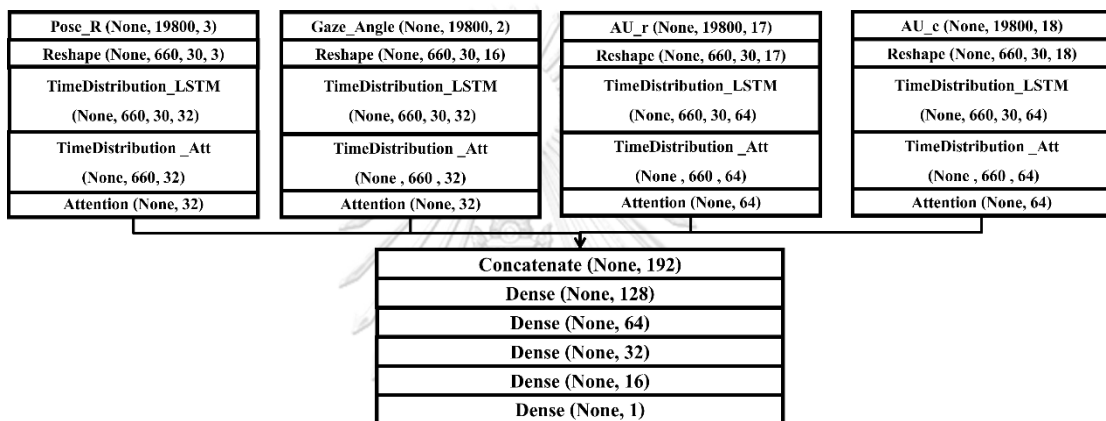


Figure 25 Intermediate Fusion Window Block LSTM Model Architecture.

#### 4.3.14 Late Fusion Window Block LSTM Model Architecture

Late Fusion Window Block LSTM Model Architecture utilizes a single window block LSTM model. Four input features are passed through their single model and averaged in the aggregation layer to determine the output. The total number of parameters is 61,068. The model architecture is shown in Figure 26.

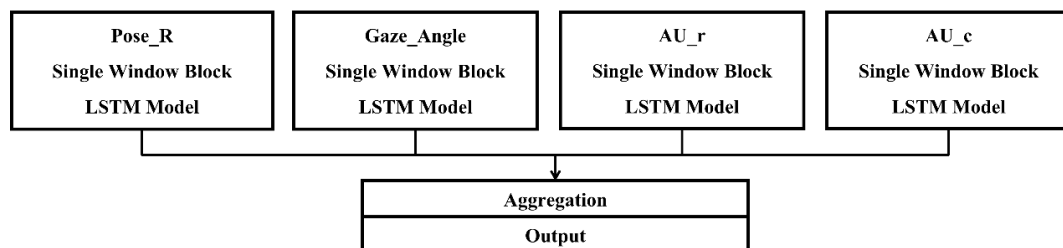


Figure 26 Late Fusion Window Block LSTM Model Architecture.

#### 4.4 Integrated Gradient Explanation

The integrated gradient is applied to the best model to explain its result. The methods to calculate the integrated gradient in time-series are the same as image classification. We compute integrated gradients for each input feature from baseline time-series input (zero-initialized time-series) to actual time-series with equally spaced intermediate steps. The integrated gradients express the contribution of their input features. Finally, we calculate the mean value of integrated gradient feature values to visualize the importance of features for depression or non-depression. We also calculate the absolute mean value of integrated gradient feature values to visualize the importance of features for arranging the important features in order.



## CHAPTER 5

### EXPERIMENTS AND RESULTS

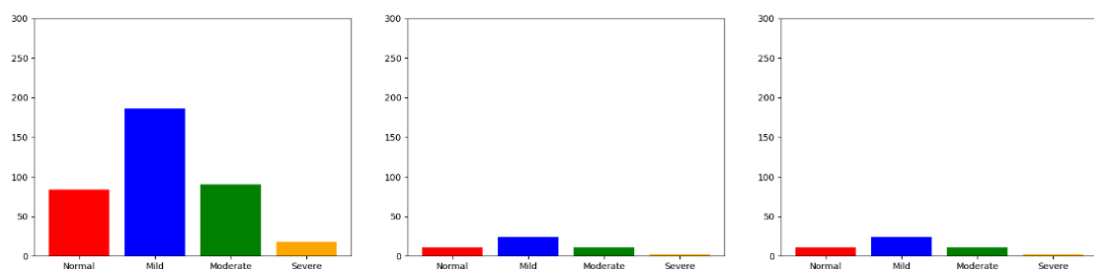
#### 5.1 Experimental Setups

##### 5.1.1 Environment Detail

A desktop computer is used to run this experiment. The processor is an Intel® Core™ i9-12900K, 12th Generation, 3.19 GHz. RAM is 64 GB. Nvidia GeForce RTX 3090 is the GPU. The operating system is Windows 10 Pro.

##### 5.1.2 Data Distribution

Raw data has 106 normal, 234 mild, 112 moderate, and 22 severe. Raw data is separated in the ratio 80:10:10 for three data sets: training, development, and testing. This three-data set is for training, validation, and testing. In 4 classes, the training data set has 84 normal, 186 mild, 90 moderate, and 18 severe. The development and testing data sets contain the same amount of data: 11 normal, 24 mild, 11 moderate, and 2 severe. Figure 27. depicts the data set with four classes. After separating, the training data set has 270 non-depressions and 108 depressions. Both the development and testing data sets contain the same amount of data, with 35 non-depressions and 13 depressions. Figure 28 depicts the data set with two classes. To balance the data, we duplicate the depression data set in the training data set shown in Figure 29. Finally, the training data set has 270 non-depressions and 216 depressions.



*Figure 27 Train Data Set, Dev Data Set, Test Data Set in 4 Class.*

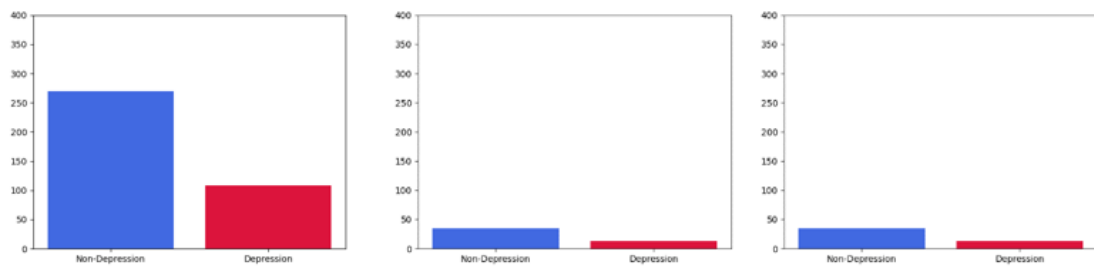


Figure 28 Train Data Set, Dev Data Set, Test Data Set in 2 Class.

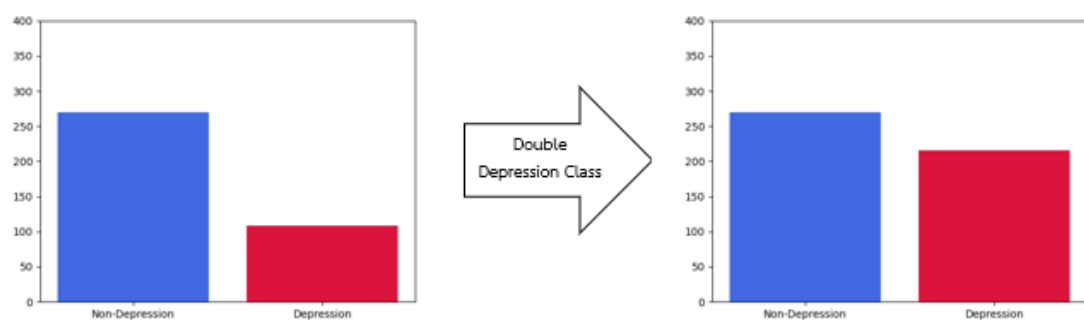


Figure 29 Double Depression Class in Train Data Set.

### 5.1.3 Implementation

The models that are used in this experiment are implemented following the Model Architecture in section 4.3 Model Architecture. After getting the results of all the models, the best model is selected to experiment with varying the label smoothing hyperparameter to improve performance.

### 5.1.4 Evaluation

The result of all models is expressed in terms of the evaluation measures described in 2.3.6 Evaluation Measures. The accuracy, confusion matrix, precision, recall, and F1 score of all models are compared to find the best model by the comparing macro F1 score.



### 5.1.5 Explanation

An integrated gradient is applied to the result of the best model to visualize the important input feature value. The important input features are expressed in terms of the impact on depression or non-depression and the order of impact.

## 5.2 Experimental Results

The result of Bi-LSTM model is shown in Table 8. As a result, action unit classification and intermediate fusion have the same accuracy of 85.42%. The difference is that intermediate fusion has better macro precision at 91.67%. Action unit classification has better macro recall at 77.91% and a better macro F1 score at 79.99%. The result of the transformer model is shown in Table 9. As a result, the best model is the intermediate fusion model, which achieves an accuracy of 83.33%, macro precision of 81.20%, macro recall of 74.07%, and macro F1 score of 76.41%. The result of the window block LSTM model is shown in Table 10. As a result, the best model is the intermediate fusion model, which achieves an accuracy of 89.58%, macro precision of 87.50%, macro recall of 85.60%, and a macro F1 score of 86.48%. The trend of almost all features between the three methods (Bi-LSTM, transformer, and window block LSTM) is similar, while window block LSTM has the best performance.

As a result, experimental models of the Bi-LSTM model, transformer model, and window block LSTM have better performance than baseline (see Table 11). The Bi-LSTM baseline achieves an accuracy of 66.78% and a macro F1 score of 40.74%. The transformer model baseline achieves an accuracy of 66.78% and a macro F1 score of 59.44%. Above the Bi-LSTM baseline, our Bi-LSTM increases to 24.25% accuracy and an 89.08% macro F1-score. The same as our transformer model, which increases a 21.21% accuracy and a 25.05% macro F1-score above the transformer baseline. The window block LSTM has the best performance, achieving an accuracy of 89.58% and a macro F1 score of 86.48%.

The intermediate fusion Bi-LSTM model and window block LSTM model were chosen to improve performance with label smoothing. The result in Table 12 shows that the intermediate fusion Bi-LSTM model with label smoothing (0.3, 0.7) achieves 91.67% accuracy, 94.87% macro precision, 84.62% macro recall, and a 88.21% macro F1-score. The result in Table 13 shows that the intermediate fusion window block LSTM model with label smoothing (0.1, 0.9) achieves

91.67% accuracy, 91.40% macro precision, 87.03% macro recall, and an 88.89% macro F1-score. Therefore, the best model is the intermediate fusion window block LSTM model with label smoothing (0.1, 0.9). Table 14 displays the predicted values of the best model for visualizing four depression levels. The predicted values show that the false positive value has only a mild level and the false negative value has only a moderate level. As a result, the model has the potential to detect normal and severe depression levels because it has more clear-cut data than mild and moderate depression levels. The accuracy of normal and severe depression levels achieves 100% in the test data set, which is useful to classify severe depression from normal people.

The integrated gradient results of the best model (intermediate fusion window block LSTM model with label smoothing) are shown in Figure 30 and Figure 31. Figure 30 shows the impact of facial features on depression or non-depression, feature by feature. Figure 31 shows the impact of overall facial features on depression or non-depression.

First, important pose features are shown in Figure 30 (A) and (B). The movement of head pose features is shown in Figure 32. The important pose features are Pose\_Rx (head nodding), Pose\_Rz (head tilting), and Pose\_Ry (head turning), respectively. Head nodding and head tilting impact non-depression, and head turning impacts depression because head nodding and head tilting are reactions of high energy and favorable to social interaction [36-39]. On the other hand, head turning means that patients have a lack of concentration on social interests and withdraw.

Second, important gaze features are shown in Figure 30 (C) and (D). The movement of eye gaze features is shown in Figure 33. The important features are Gaze\_y (looking up or down) and Gaze\_x (looking left or right), respectively. Both gaze features impact depression because looking around, having a nonspecific gaze, and not having eye contact mean patients have a lack of concentration and are absent-minded [36-38]. The reduction in eye movement is justified as a depressive symptom [28].

Third, important action unit regression features are shown in Figure 30 (E) and (F). The movement of the action unit is shown in Table 1. The obvious features that impact depression are the AU26 jaw drop, AU20 lip stretcher, and AU07 lid tightener, which represent grumbling, frowning, and scowling faces that relate to negative feelings and social disinterest [38]. In controversy, the features that impact non-depression are the AU06 cheek raiser, AU25 lips part, AU14 dimpler, and AU12 lip corner puller, which represent the posture of talking and smiling.

Forth, action unit classification important features are shown in Figure 30 (G) and (H). The movement of the action unit is shown in Table 1. The obvious features that impact depression are the AU07 lid tightener, the AU26 jaw drop, and the AU25 lips part. In the same direction as action unit regression, AU07 and AU26 represent grumbling, frowning, and scowling faces that relate to negative feelings and social disinterest. However, AU25 represents when people talk. In the same direction as [38], silence and speaking can be justified as depression or non-depression depending on the speech content. On the other hand, the features that impact non-depression are the AU23 lip tightener, the AU12 lip corner puller, the AU45 blink, and the AU09 nose wrinkle. They represent pursing lips, smiling, and blinking.

Finally, overall features are shown in Figure 31 (A) and (B). The important features are action unit classification, action unit regression, head pose, and gaze, respectively. Facial expression can be detected via action unit classification and action unit regression, which make it easy to observe depression like human observation [27]. Concentration and social interest can be detected via head pose and gaze. Therefore, machine learning can detect depression through four main features in the same ways as human observation.

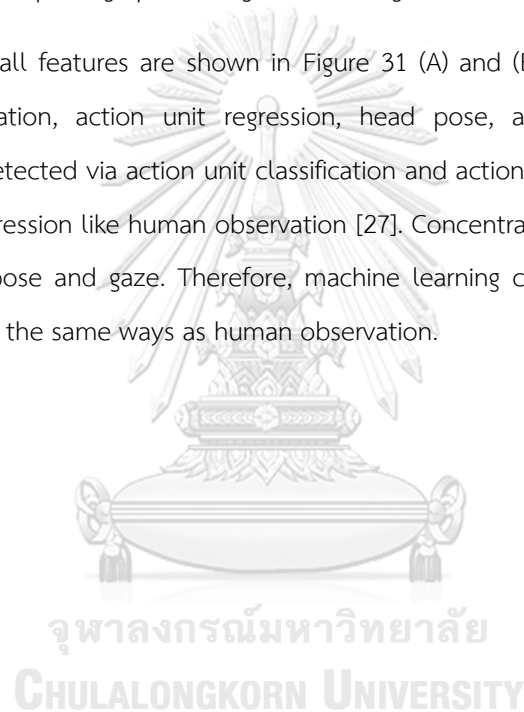


Table 8 Bi-LSTM Model Result. Hightlighted numbers refer to the winners.

Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
Pose	6	7	32	3	79.17	74.36	68.79	70.52	77.88	79.17	77.84
Gaze	8	5	31	4	81.25	76.39	75.05	75.66	80.84	81.25	81.01
AU_r	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27
AU_c	8	5	33	2	85.42	83.42	77.91	79.99	84.99	85.42	84.77
Early Fusion	6	7	32	3	79.17	74.36	68.79	70.52	77.88	79.17	77.84
Intermediate Fusion	6	7	35	0	85.42	91.67	73.08	77.03	87.85	85.42	83.39
Late Fusion	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27

Table 9 Transformer Model Result. Highlighted numbers refer to the winners.

Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
Pose	6	7	32	3	79.17	74.36	68.79	70.52	77.88	79.17	77.84
Gaze	0	13	35	0	72.92	36.46	50.00	42.17	53.17	72.92	61.50
AU_r	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
AU_c	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
Early Fusion	5	8	32	3	77.08	71.25	64.95	66.48	75.26	77.08	75.12
Intermediate Fusion	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27
Late Fusion	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41

Table 10 Window Block LSTM Model Result. Highlighted numbers refer to the winners.

Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)			
						Precision	Recall	F1	Precision	Recall	F1	
Window Block LSTM	Pose	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
	Gaze	8	5	33	2	85.42	83.42	77.91	79.99	84.99	85.42	84.77
	AU_r	6	7	34	1	83.33	84.32	71.65	74.74	83.68	83.33	81.49
	AU_c	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
	Early Fusion	8	5	31	4	81.25	76.39	75.05	75.66	80.84	81.25	81.01
	Intermediate Fusion	10	3	33	2	89.58	87.50	85.60	86.48	89.41	89.58	89.45
Late Fusion	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27	

Table 11 Baseline Comparison. Highlighted numbers refer to the winners.

Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
Baseline Bi-LSTM	0	13	33	2	68.75	35.87	47.14	40.74	52.31	68.75	59.41
Baseline Transformer	5	8	28	7	68.75	59.72	59.23	59.44	68.00	68.75	68.35
Our Bi-LSTM	6	7	35	0	85.42	91.67	73.08	77.03	87.85	85.42	83.39
Our Transformer	7	6	33	2	83.33	81.20	74.07	76.41	82.76	83.33	82.27
Our Window Block LSTM	10	3	33	2	89.58	87.50	85.60	86.48	89.41	89.58	89.45

Table 12 Intermediate Fusion Bi-LSTM Model with Label Smoothing Result. Highlighted numbers refer to the winners.

Intermediate Fusion Bi-LSTM Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
(0, 1)	6	7	35	0	0.8542	0.9167	0.7308	0.7703	0.8785	0.8542	0.8339
(0.05, 0.95)	5	8	33	2	0.7917	0.7596	0.6637	0.6842	0.7803	0.7917	0.7686
(0.1, 0.9)	6	7	33	2	0.8125	0.7875	0.7022	0.7257	0.8047	0.8125	0.7964
(0.15, 0.85)	7	6	33	2	0.8333	0.8120	0.7407	0.7641	0.8276	0.8333	0.8227
(0.2, 0.8)	6	7	32	3	0.7917	0.7436	0.6879	0.7052	0.7788	0.7917	0.7784
(0.25, 0.75)	5	8	34	1	0.8125	0.8214	0.6780	0.7047	0.8160	0.8125	0.7865
(0.3, 0.7)	9	4	35	0	0.9167	0.9487	0.8462	0.8821	0.9252	0.9167	0.9113
(0.35, 0.65)	5	8	34	1	0.8125	0.8214	0.6780	0.7047	0.8160	0.8125	0.7865
(0.4, 0.6)	5	8	34	1	0.8125	0.8214	0.6780	0.7047	0.8160	0.8125	0.7865
(0.45, 0.55)	6	7	34	1	0.8333	0.8432	0.7165	0.7474	0.8368	0.8333	0.8149

Table 13 Intermediate Fusion Window Block LSTM Model with Label Smoothing Result. Highlighted numbers refer to the winners.

Intermediate Fusion Window Block LSTM Model	TP	FN	TN	FP	Accuracy (%)	Macro (%)			Micro (%)		
						Precision	Recall	F1	Precision	Recall	F1
(0, 1)	10	3	33	2	89.58	87.50	85.60	86.48	89.41	89.58	89.45
(0.05, 0.95)	10	3	34	1	91.67	91.40	87.03	88.89	91.63	91.67	91.44
(0.1, 0.9)	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
(0.15, 0.85)	6	7	33	2	81.25	78.75	70.22	72.57	80.47	81.25	79.64
(0.2, 0.8)	4	9	34	1	79.17	79.53	63.96	65.81	79.32	79.17	75.61
(0.25, 0.75)	6	7	32	3	79.17	74.36	68.79	70.52	77.88	79.17	77.84
(0.3, 0.7)	6	7	34	1	83.33	84.32	71.65	74.74	83.68	83.33	81.49
(0.35, 0.65)	6	7	33	2	81.25	78.75	70.22	72.57	80.47	81.25	79.64
(0.4, 0.6)	7	6	32	3	81.25	77.11	72.64	74.27	80.36	81.25	80.41
(0.45, 0.55)	6	7	34	1	85.42	86.25	75.49	78.67	85.68	85.42	84.17



Table 14 Predicted Values of Intermediate Fusion Window Block LSTM Model with Label Smoothing (0.05, 0.95)

Predicted Values	Normal	Mild	Moderate	Severe
True Positive	-	-	8	2
False Negative	-	-	3	0
True Negative	11	23	-	-
False Positive	0	1	-	-



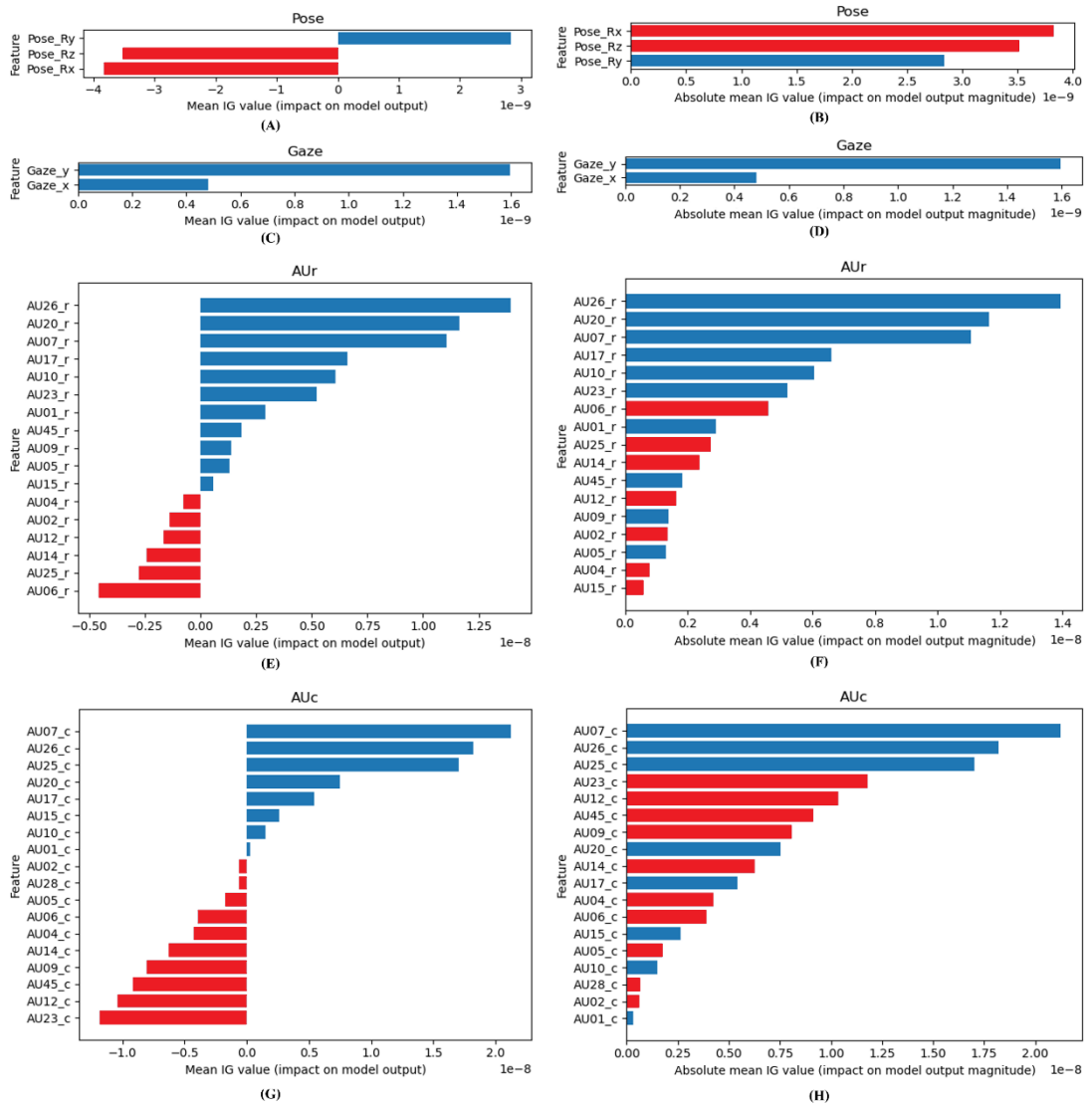


Figure 30 (A) Pose impact on model output, (B) Pose impact on model output magnitude, (C) Gaze impact on model output, (D) Gaze impact on model output magnitude, (E) AUr impact on model output, (F) AUr impact on model output magnitude, (G) AUc impact on model output, (H) AUc impact on model output magnitude.

\* Red color refers to a negative effect (tends to be non-depressive)

\*\* Blue color refers to a positive effect (tends to be depressive).

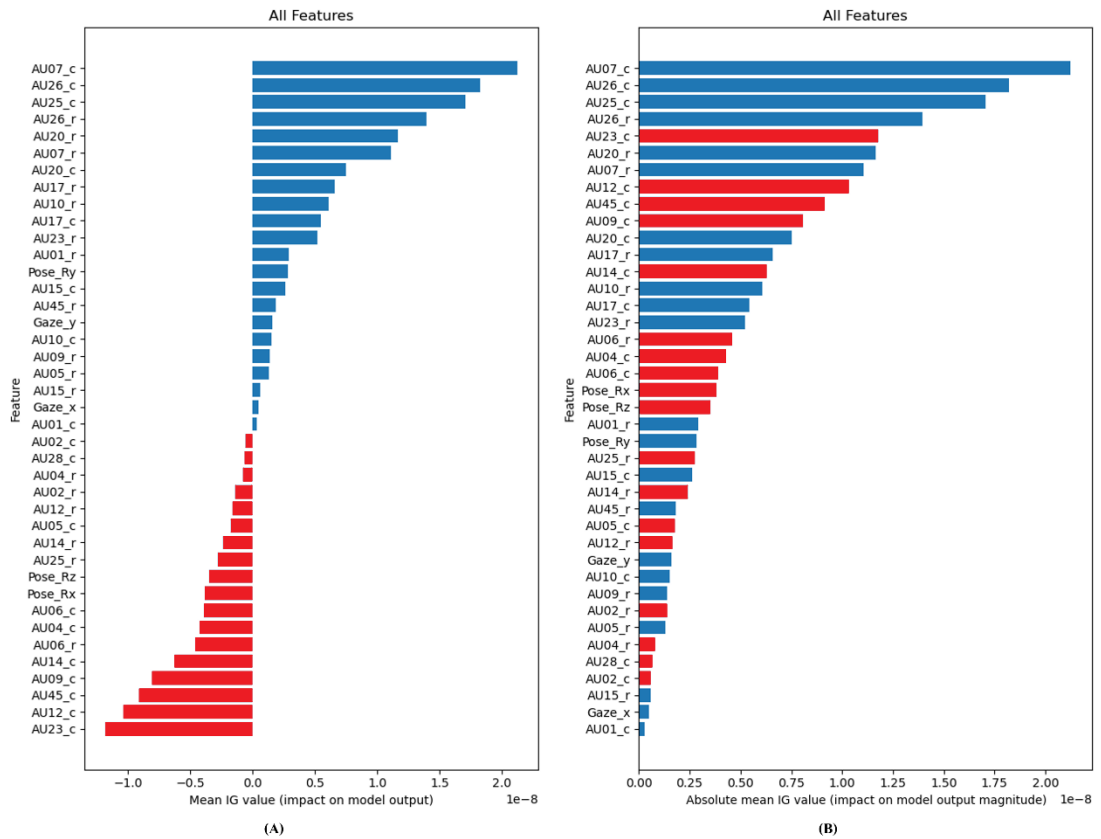


Figure 31 (A) Positive/negative impact of all features, (B) Absolute impact (magnitude) of all features.

\* Red color refers to a negative effect (tends to be non-depressive).

\*\* Blue color refers to a positive effect (tends to be depressive).

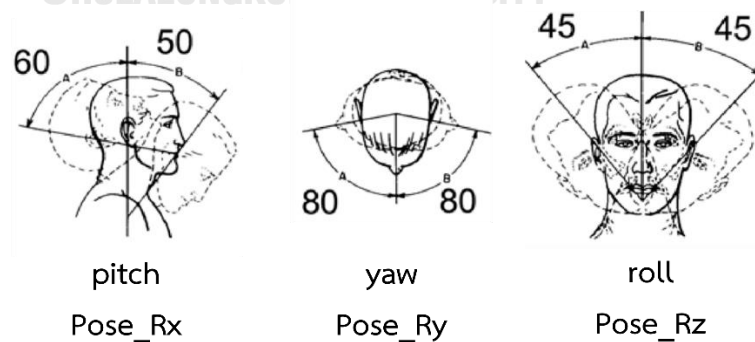


Figure 32 Head Pose Movement [40].

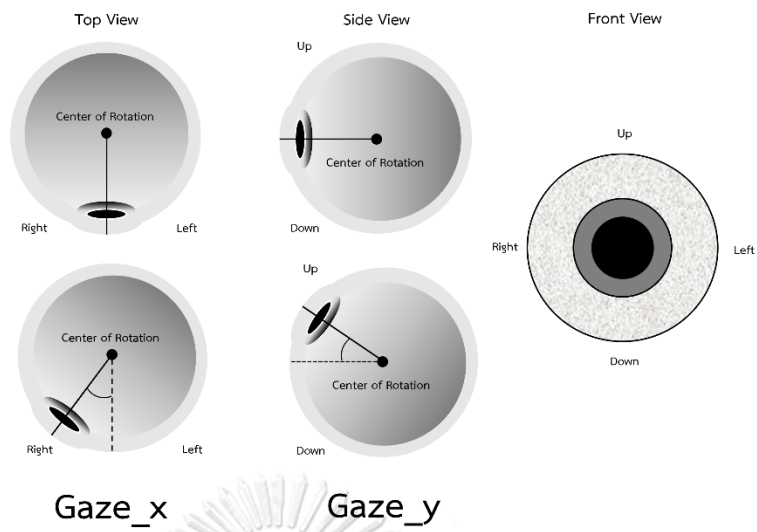


Figure 33 Gaze Movement.

## CHAPTER 6

### CONCLUSION AND FUTURE WORK

#### Conclusion

Machine learning models can detect depression from extracted facial features in time-series format without utilizing the original interview video to protect privacy. The well-known techniques of machine learning that were experimented with in this thesis are the Bi-LSTM model, transformer model, and window block LSTM model. All our experimental models have better performance than both the Bi-LSTM baseline and the transformer baseline. Three types of fusion methods—early fusion, intermediate fusion, and late fusion—are applied to our model. We also applied label smoothing to improve performance. The best model is intermediate fusion window block LSTM with label smoothing (0.05, 0.95), which achieves 91.67% accuracy, 91.40% macro precision, 87.03% macro recall, and 88.89% macro F1-score.

The important key features that have an influence on depression detection are action unit classification, action unit regression, pose rotation, and gaze angle, respectively. All features indicate that patients who have depression symptoms keep frowning, grumbling, scowling, head turning, no specific gaze, and slow eye movement, which express a lack of concentration, social disinterest, and negative feelings.

#### Future Work

The label smoothing techniques can be applied in several ways to set up experiments to improve model performance since the original depression classes are four and the extracted features from the Openface tool do not have 100% accuracy. We can apply different label smoothing values for normal, mild, moderate, and severe classes for binary classification. In the same direction, extracted features that have poor accuracy can utilize label smoothing techniques to prevent a model from becoming overconfident in its predictions.

## REFERENCES

1. Calegario, V.C., et al., *Closed doors: Predictors of stress, anxiety, depression, and PTSD during the onset of COVID-19 pandemic in Brazil*. *Journal of affective disorders*, 2022. **310**: p. 441-451.
2. Cheng, X., et al., *Prevalence of depressive disorders and associated demographic characteristics in Shandong: An epidemiological investigation*. *Journal of Affective Disorders*, 2022. **311**: p. 198-204.
3. Hawes, M.T., et al., *Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic*. *Psychological medicine*, 2022. **52**(14): p. 3222-3230.
4. Santomauro, D.F., et al., *Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic*. *The Lancet*, 2021. **398**(10312): p. 1700-1712.
5. Tabur, A., et al. *Anxiety, burnout and depression, psychological well-being as predictor of healthcare professionals' turnover during the COVID-19 pandemic: study in a pandemic hospital*. in *Healthcare*. 2022. MDPI.
6. Smith, K.M., P.F. Renshaw, and J. Bilello, *The diagnosis of depression: current and emerging methods*. *Comprehensive psychiatry*, 2013. **54**(1): p. 1-6.
7. Paul Ekman, Wallace V. Friesen, and J.C. Hager., *Facial Action Coding System*. 2002, 545 East 4500 South E-160, Salt Lake City UT 84107: Research Nexus division of Network Information Research Corporation.
8. Baltrusaitis, T., et al. *Openface 2.0: Facial behavior analysis toolkit*. in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. 2018. IEEE.
9. Zadeh, A., et al. *Convolutional experts constrained local model for 3d facial landmark detection*. in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017.
10. Baltrusaitis, T., P. Robinson, and L.-P. Morency. *Constrained local neural fields for robust facial landmark detection in the wild*. in *Proceedings of the IEEE international conference on computer vision workshops*. 2013.

11. Wood, E., et al. *Rendering of eyes for eye-shape registration and gaze estimation*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
12. Baltrušaitis, T., M. Mahmoud, and P. Robinson. *Cross-dataset learning and person-specific normalisation for automatic action unit detection*. in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. 2015. IEEE.
13. La Cascia, M., S. Sclaroff, and V. Athitsos, *Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models*. IEEE Transactions on pattern analysis and machine intelligence, 2000. **22**(4): p. 322-336.
14. Baltrušaitis, T., P. Robinson, and L.-P. Morency. *3D constrained local model for rigid and non-rigid facial tracking*. in *2012 IEEE conference on computer vision and pattern recognition*. 2012. IEEE.
15. Zhang, X., et al. *Appearance-based gaze estimation in the wild*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.
16. Mavadati, S.M., et al., *Disfa: A spontaneous facial action intensity database*. IEEE Transactions on Affective Computing, 2013. **4**(2): p. 151-160.
17. Huang, S.-C., et al., *Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines*. NPJ digital medicine, 2020. **3**(1): p. 136.
18. Hochreiter, S. and J. Schmidhuber, *Long short-term memory*. Neural computation, 1997. **9**(8): p. 1735-1780.
19. Galstyan, A. and P.R. Cohen. *Empirical comparison of “hard” and “soft” label propagation for relational classification*. in *International Conference on Inductive Logic Programming*. 2007. Springer.
20. Bahdanau, D., K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473, 2014.
21. Vaswani, A., et al., *Attention is all you need*. Advances in neural information processing systems, 2017. **30**.
22. Sundararajan, M., A. Taly, and Q. Yan. *Axiomatic attribution for deep networks*. in *International conference on machine learning*. 2017. PMLR.
23. Watanabe, A. and T. Yamazaki, *Representation of the brain network by*

- electroencephalograms during facial expressions*. Journal of Neuroscience Methods, 2021. **357**: p. 109158.
24. Fu, C.H., et al., *Neural responses to happy facial expressions in major depression following antidepressant treatment*. American Journal of Psychiatry, 2007. **164**(4): p. 599-607.
  25. Nakamura, A., et al., *The cerebellum as a moderator of negative bias of facial expression processing in depressive patients*. Journal of Affective Disorders Reports, 2022. **7**: p. 100295.
  26. Schirmer, A. and R. Adolphs, *Emotion perception from face, voice, and touch: comparisons and convergence*. Trends in cognitive sciences, 2017. **21**(3): p. 216-228.
  27. Scott, N.J., et al., *Facial cues to depressive symptoms and their associated personality attributions*. Psychiatry Research, 2013. **208**(1): p. 47-53.
  28. Stolicyn, A., J.D. Steele, and P. Seriès, *Prediction of depression symptoms in individual subjects with face and eye movement tracking*. Psychological medicine, 2022. **52**(9): p. 1784-1792.
  29. Liu, G.-D., et al., *A brief review of artificial intelligence applications and algorithms for psychiatric disorders*. Engineering, 2020. **6**(4): p. 462-467.
  30. Nasser, S.A., I.A. Hashim, and W.H. Ali. *A review on depression detection and diagnoses based on visual facial cues*. in *2020 3rd International Conference on Engineering Technology and its Applications (IICETA)*. 2020. IEEE.
  31. Ray, A., et al. *Multi-level attention network using text, audio and video for depression prediction*. in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*. 2019.
  32. Akbar, H., et al. *Exploiting facial action unit in video for recognizing depression using metaheuristic and neural networks*. in *2021 1st International conference on computer science and artificial intelligence (ICCSAI)*. 2021. IEEE.
  33. Sun, H., et al., *Multi-modal adaptive fusion transformer network for the estimation of depression level*. Sensors, 2021. **21**(14): p. 4764.
  34. Rathi, S., B. Kaur, and R. Agrawal, *Selection of relevant visual feature sets for enhanced depression detection using incremental linear discriminant analysis*. Multimedia Tools and Applications, 2022. **81**(13): p. 17703-17727.



35. Schultebrucks, K., et al., *Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood*. *Psychological Medicine*, 2022. **52**(5): p. 957-967.
36. Fossi, L., C. Faravelli, and M. Paoli, *The ethological approach to the assessment of depressive disorders*. *The Journal of nervous and mental disease*, 1984. **172**(6): p. 332-341.
37. Schelde, J.T.M., *Major depression: Behavioral markers of depression and recovery*. *The Journal of nervous and mental disease*, 1998. **186**(3): p. 133-140.
38. Fiquer, J.T., P.S. Boggio, and C. Gorenstein, *Talking bodies: nonverbal behavior in the assessment of depression severity*. *Journal of affective disorders*, 2013. **150**(3): p. 1114-1119.
39. Gahalawat, M., et al. *Explainable Depression Detection via Head Motion Patterns*. in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023.
40. Doss, A.S.A., et al., *A comprehensive review of wearable assistive robotic devices used for head and neck rehabilitation*. *Results in Engineering*, 2023. **19**: p. 101306.



จุฬาลงกรณ์มหาวิทยาลัย  
**CHULALONGKORN UNIVERSITY**

## VITA

NAME Yanisa Mahayossanunt

DATE OF BIRTH 11 July 1997

PLACE OF BIRTH Bangkok, Thailand

INSTITUTIONS ATTENDED B.Eng., Thai-Nichi Institute of Technology  
M.Eng., Chulalongkorn University

