Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2025, 2025:134
https://doi.org/10.28919/cmbn/9590
ISSN: 2052-2541

DEPRESSION SEVERITY DETECTION FROM FACIAL EXPRESSIONS IN VIDEOS USING RECURRENT NEURAL NETWORKS

BRILYAN NATHANAEL RUMAHORBO[1,*], GREGORIUS NATANAEL ELWIREHARDJA[2,3], BENS PARDAMEAN[1,2]

[1]Computer Science Department BINUS Graduate Program – Master of Computer Science Program Bina Nusantara University, Jakarta 11480, Indonesia

[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia

[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

Abstract. Major Depressive Disorder (MDD), commonly known as depression, impacts over 300 million people worldwide. Diagnosing this condition often relies on subjective judgment, emphasizing the need for an objective approach. Deep learning, particularly Recurrent Neural Networks (RNNs), offers a promising solution, especially for analyzing multivariate time series data. Researchers have explored the use of RNNs and their variants to detect depression severity. Therefore, this study conducts experimental testing to identify depression severity using first derivative techniques and feature engineering on the RNN model and its variations. The first derivative is calculated for each frame during the subject's interview, and feature engineering techniques focus on the eyes and lips, known to be associated with depression, including calculations of upper and lower lip distances, eye openness, and more. The RNN model, incorporating feature engineering, achieved the best results among the proposed methods, with a Mean Absolute Error (MAE) of 5.04 and Root Mean Squared Error (RMSE) of 6.03. However, further performance enhancement is possible by increasing the number of layers and neurons, considering the current model's relative simplicity due to limited resources.

[*]Corresponding author
E-mail address: brilyan.rumahorbo@binus.ac.id

Received September 06, 2025

Keywords: major depressive disorder; deep learning; RNN; LSTM; GRU; facial expression.

2020 AMS Subject Classification: 68T07, 68T01, 26A24.

1. INTRODUCTION

Major Depressive Disorder (MDD), also known as depression, is a mental disorder that affects individuals psychologically [1]. Those experiencing this disorder often endure persistent feelings of sadness, loss of interest or pleasure in once-enjoyable activities, and even thoughts of suicide [2]. According to Yuan et al. [3], MDD is a prevalent global issue, with over 300 million people having experienced depressive disorders. Moreover, Jiang et al. [4] also indicates that approximately one million individuals worldwide commit suicide each year due to MDD. The suicide rate in the United States and China is approximately 85.3 per 100,000 people and 100 per 100,000 people, respectively, due to MDD.

According to Jiang et al. [4], the rising suicide rate is attributed to patients not receiving prompt and accurate diagnosis and treatment. The diagnostic methods employed for detecting depression have been proven to be unreliable and subjective [5]. Diagnosing mental disorders based on an individual's behavior is challenging and time-consuming. Doctors require more time for diagnosing depressive disorders, leading to potential variations in diagnoses based on each doctor's experience [6]. The Patient Health Questionnaire (PHQ) serves as a screening tool, posing questions about the frequency of experiencing depression symptoms. The PHQ score ranges from 0 to 24, with higher scores indicating more severe depression symptoms [7].

To enhance the accuracy of depression assessment, a technology known as Computer-Aided Detection (CAD) is employed [8]. This technology utilizes neuroimaging data, offering more objective results for diagnosing mental disorders [9]. Researchers have also tried to develop deep learning (DL) technologies [10], such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to help doctors diagnose mental disorders [11]. Common methods for diagnosing mental disorders often involve assessing body temperature, heart rate, respiratory rate, and blood pressure [12]. However, this research takes a different approach by discussing the utilization of RNN models and their variants, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), for detecting depression based on facial expressions. RNNs are a popular type of DL model with effective variants [13], and this study specifically employs an LSTM model, which has been proven effective in learning and modeling sequential data [14]. In previous research, considered as the baseline for this study, Rasipuram et al. [15] employed the LSTM model (Long Short-Term Memory), a variant of RNN, achieving an MAE (Mean Absolute Error) of 4.83 and an RMSE (Root Mean Square Error) of 5.76 for depression detection. However, this study utilized multimodality, and the results could potentially be more optimal if appropriate features were utilized.

This study evaluated the ability of two methods to detect depression based on facial expressions, which used first derivative data and feature engineering data, respectively. In the first derivative approach, the model learned from the changes in facial expressions between successive frames. This method was inspired by Rumahorbo et al. [1] that used the XGBoost model on first derivative data to achieve 6.22 RMSE, which led to the development of this method served as the baseline in this study. However, the study did not divide time segments, leading to inconsistent training data lengths and making it hard for the model to learn.

In the feature engineering approach, the research focuses on the eyes and lips because these regions can be affected by depression. This study also uses the concept of feature engineering inspired by Thati et al. [16], which used feature engineering and a machine learning model called Support Vector Machine (SVM) to achieve an accuracy of 83%. Thus, a novelty of this study is shown in it uses feature engineering to extract specific features from facial expressions, such as eye openness, upper and lower lip distances, lip length and height, and then uses RNN models and their variants to detect depression based on these features.

Overall, we chose this method based on a previous systematic literature review which is related to the chosen method [17]. It is expected that the outcome of this research to assist doctors in diagnosing depression more objectively and accurately. Furthermore, this research may open opportunities for future researchers by providing new insights into identifying depression based on facial expressions in videos. However, a significant challenge in this study is that employing an excessive number of layers in the architecture may result in suboptimal model performance and hinder its ability to adapt effectively to new datasets [18]. To address this challenge, careful selection of layers and fine-tuning strategies are essential to balance model complexity with generalization, ensuring robust performance across diverse data.

2. Related Works

The research carried out by Thati et al. [16] applied complex feature engineering techniques, including measurement of lip angles, upper lip raising, lip tilt, nose wrinkles, and others, to analyze depression of an individual. This study used various Machine Learning (ML) methods and demonstrated that SVM achieved an accuracy of 83%. In another study, Rumahorbo et al. [1] used different ML models, such as XGBoost, for depression detection and obtained MAE and RMSE scores of 5.28 and 6.22, respectively.

Additionally, other researchers have employed Deep Learning (DL) models to detect depression. Rasipuram et al. [15] utilized a variant model, namely LSTM. The study implemented multimodality, incorporating video, audio, and text, and achieved excellent results with 4.83 MAE and 5.76 RMSE. Furthermore, Muzammel et al. [19] compared deep learning models for depression analysis, with the LSTM + MFCC combination outperforming others with an F1-Score of 0.85.

Flores et al. [20] and Ceccarelli et al. [21] provided new insights, indicating a strong correlation between eye gaze and depression. This correlation is substantiated by their research results, achieving an F1-Score of 0.93 and exhibiting an accuracy level of 0.7489. Additionally, the study by Cao et al. [22] introduced a unique aspect by employing a different model variant, namely GRU, and achieving an accuracy of 89.77% using only Facial Action Unit (FAU) features.

Overall, previous researchers have made significant contributions in their studies. Both LSTM and GRU have demonstrated strong performance in depression analysis. Additionally, Thati et al. [16] has opened up the potential for further studies in analyzing depression using feature engineering techniques on facial expressions in videos. The study has paved the way for this research to experiment with feature engineering techniques, serving as a novelty in this study. This research will specifically focus on delving deeper into facial feature engineering and its application in deep learning with RNN models and their variations.

3. METHODOLOGY

3.1. Architecture. In this study, the model was trained using multivariate time series data and was fine-tuned through Grid Search. Grid search experiments were conducted to evaluate the following hyperparameter combinations: the number of neurons (16, 32, 64), hidden layers (1, 2), learning rates (0.01 and 0.001), batch size (16, 32), dropout rate (0.5), and epochs (10). Based on the experiments conducted, the optimal hyperparameter tuning for the first derivative was presented in Table 1, and for feature engineering, was shown in Table 2. The architectural overview of the model design for the first derivative and the engineering of the features is illustrated in Figure 1.

TABLE 1. Best hyperparameter tuning results for first derivative

Hyperparameter	RNN	LSTM	GRU
Learning rate	0.01	0.01	0.01
Hidden layers	2	2	2
Batch size	32	32	32
Number of neurons	64	32	32
Dropout	0.5	0.5	0.5

TABLE 2. Best hyperparameter tuning results for feature engineering

Hyperparameter	RNN	LSTM	GRU
Learning rate	0.01	0.01	0.01
Hidden layers	2	2	2
Batch size	32	32	32
Number of neurons	64	64	64
Dropout	0.5	0.5	0.5

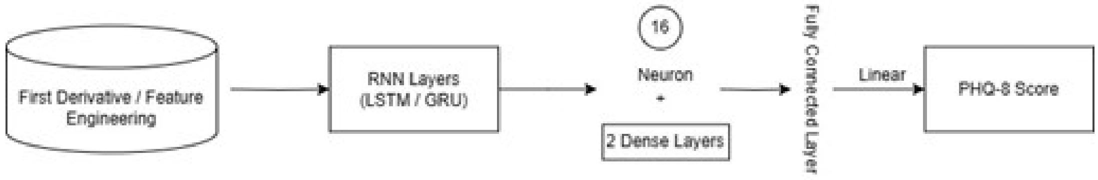


FIGURE 1. Architecture of the model

3.2. Dataset. This study used the publicly available Distress Analysis Interview Corpus (DAIC-WOZ) dataset [23], which comprised interview data recorded in various formats, including audio, video, and text [24]. The DAIC-WOZ dataset was widely recognized for its large size and frequent use in mental health research [25]. To distinguish between individuals with normal mental health and those experiencing severe depression, the study relied on Patient Health Questionnaire-8 (PHQ-8) scores. A PHQ-8 score of 10 indicated depression, while a score of PHQ-8 $<$ 10 was considered normal [26]. Therefore, the objective of this study was to detect the severity of an individual's depression based on their PHQ-8 scores. Table 3 provides information about the DAIC-WOZ dataset.

TABLE 3. Label distribution of the subjects in the DAIC-WOZ dataset

Subset	Depression	Normal	Total
Train	10	77	107
Development	15	23	35
Test	20	33	47

In this dataset, three modalities (audio, text, and video) were present. However, this study specifically focused on the text and video modalities. Text was utilized for interview conversation transcripts, while the video modality encompassed Facial Action Unit (FAU), feature3D, and eye gaze data.

Facial Action Unit (FAU) is a set of points on a person's face used to identify specific facial expressions and understand the emotional meaning conveyed by those expressions [27]. Features3D comprises 68 landmark points on the face, each with X, Y, and Z coordinate lines. Gaze pertains to how someone looks or gazes during the interview process.

3.3. Data Preprocessing. Based on Figure 2, the researcher elaborated in detail the workflow of data preprocessing conducted in this study. In the first stage, the process began with acquiring data from the DAIC-WOZ dataset. This dataset comprised multivariate time series data that could be used to identify an individual's mental health. Multivariate time series data enabled data manipulation, such as conducting statistical calculations.

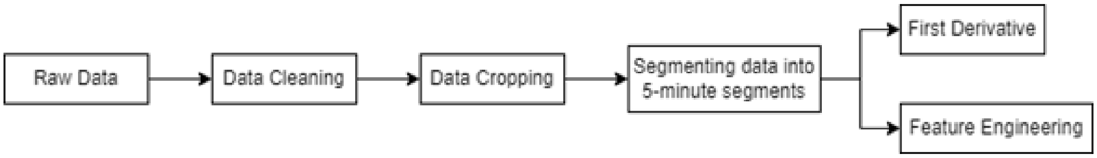
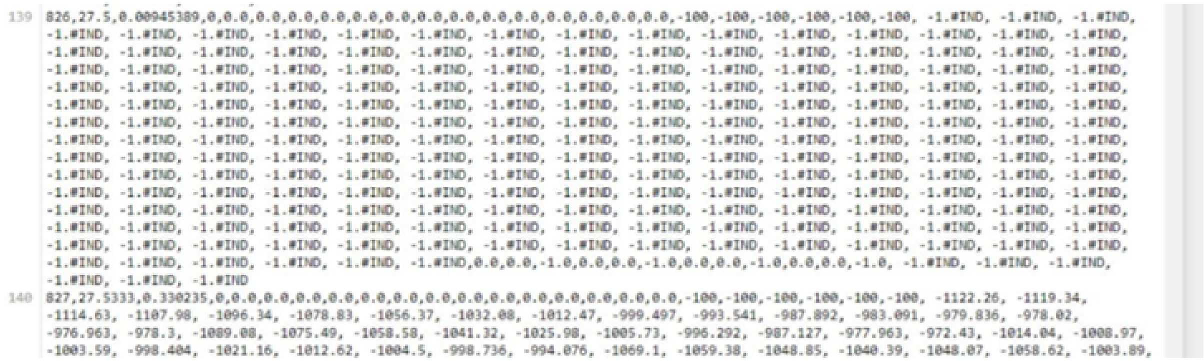


FIGURE 2. Workflow of Data Preprocessing

In the second stage, which focused on data cleaning, the goal was to identify and address missing data, outliers, and errors to ensure accuracy and reliability. For instance, in Figure 3, anomalies were observed in the data of patient number 432. Certain parts were poorly recorded, leading to the entry '#IND', which made the data invalid. Therefore, it was essential to remove such data segments to achieve optimal results.



```
826,27.5,0.00945389,0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,-100,-100,-100,-100,-100,-100, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND, -1.#IND,0.0,0.0,-1.0,0.0,0.0,-1.0,0.0,0.0,-1.0,0.0,0.0,-1.0, -1.#IND, -1.#IND, -1.#IND,
-1.#IND, -1.#IND, -1.#IND
827,27.5333,0.330235,0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,-100,-100,-100,-100,-100,-100, -1122.26, -1119.34,
-1114.63, -1107.98, -1096.34, -1078.83, -1056.37, -1032.08, -1012.47, -999.497, -993.541, -987.892, -983.091, -979.836, -978.02,
-976.963, -978.3, -1089.08, -1075.49, -1058.58, -1041.32, -1025.98, -1005.73, -996.292, -987.127, -977.963, -972.43, -1014.04, -1008.97,
-1003.59, -998.404, -1021.16, -1012.62, -1004.5, -998.736, -994.076, -1069.1, -1059.38, -1048.85, -1040.39, -1048.07, -1058.62, -1003.89,
```

FIGURE 3. The anomaly in the data of patient number 432

Next, the third stage was data cropping. This stage involved trimming the interview duration based on timestamps (start_time, end_time), specifically when the subject entered the video frame and initiated interaction or concluded the interview. The purpose of this trimming was to reduce the complexity of the data processed by the model and to focus the analysis on the specified time range.

TABLE 4. Data segmentation rules used in this experiment

Segment	Time	Frames
1	0-5 minutes	1-9000
2	5-10 minutes	9001-18001
3	10-15 minutes	18002-27002
4	15-18 minutes	disergarded

In the fourth stage, the researcher proceeded with segmenting the data into several segments for each patient, with each segment lasting 5 minutes. This was done to increase the amount of data and facilitate the model learning process, ensuring consistent time durations within the data and with the aim of enhancing learning performance. To perform data segmentation, it required 9000 frames or 300 timestamps. However, if the data division in one segment did not reach more than 5 minutes, that segment was disregarded. For example, if there was video data with a duration of 18 minutes, the data segmentation was conducted as shown in Table 4.

In the final stage, the researcher conducted the division of previously preprocessed data. This division covered the proposed methods, namely the first derivative and feature engineering, which were novelties in this study. The division was carried out by duplicating the data and categorizing it based on the proposed methods for further stages.

3.3.1. *First Derivative.* The first technique, known as the first derivative, referred to measuring changes in intensity or differences in pixels in facial images. In the context of depression detection, changes or variations in facial expressions could serve as indicators. This method involved monitoring changes in facial expressions. In this study, the first derivative method was applied to each segment in the patient data, allowing for specific change measurements. This first derivative entailed the difference between each row and its preceding row, and the results were exported for each segment.

TABLE 5. The p-value of correlated features

Features	p-value
AU12_r	0.012
AU12_c	0.061
y_h1	0.061

Next, the researcher conducted feature selection based on correlation using Pearson Correlation (PC). Researchers typically used a threshold of p-value ≤ 0.05, but in this first derivative method, only one feature was obtained using that threshold. Therefore, the researcher slightly increased the threshold and used a p-value ≤ 0.065. The results of feature selection indicated that the eye and lip areas had significant features, including AU12_r, AU12_c, and y_h1, as presented in Table 5.

3.3.2. *Feature Engineering.*

The feature engineering technique included feature selection and manipulated features as part of this stage. Feature engineering on the face involved manipulating and adjusting facial expression features to improve the quality of information for the model. This included feature extraction, such as lip and eye shape.

Based on the results of the previous research data collection, the eyes and lips were identified as the most significant facial areas in detecting someone's depression. Therefore, the researcher plotted points on the individual's face during interview sessions, aiding in the specific feature selection for the eyes and lips area. In Figure 4, the areas of the eyes and lips are clearly visible. The researcher extracted features at landmarks 36–41 for the left eye, 42–47 for the right eye, and 48–67 for the lips.

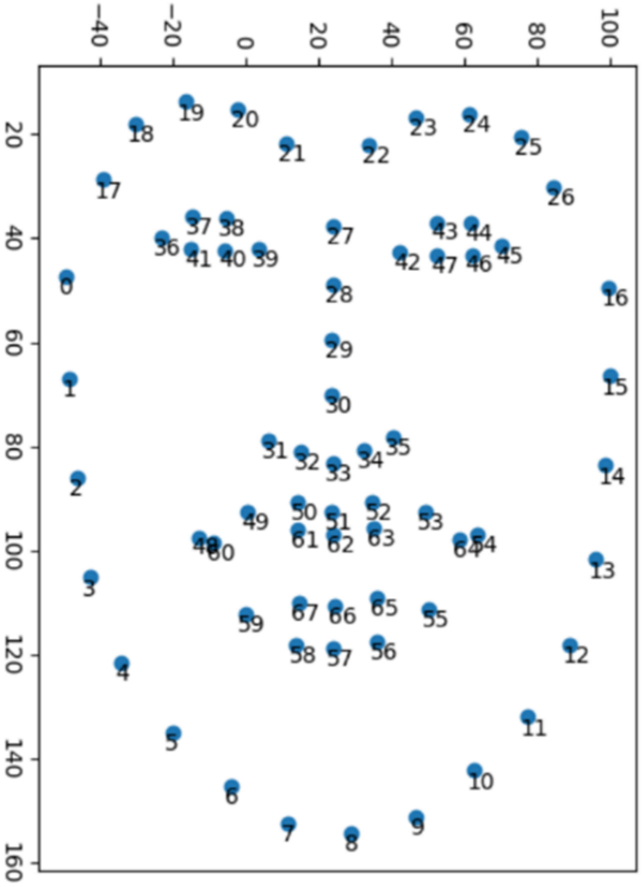


FIGURE 4. Landmark points on the face of the subject

The feature engineering applied in this study involved calculating the length and height of the lips, the distance between the lower and upper lips, and measuring eye openness for both the right and left eyes. In this context, these features were processed intricately to enhance the model's ability to detect depression based on the facial expressions of patients. Detailed information on the calculation of manipulated features is shown in Table 6.

TABLE 6. Feature engineering processing

Features	Feature engineering processing technique
width_lips_r	abs(57-48)
height_lips_c	abs(57-51)
left_eye_openness	abs(41-37)
right_eye_openness	abs(46-44)
upper_lower_lips	abs(66-62)

$$\text{Distance} = |X_1 - X_2| \tag{1}$$

The formula shown above was used for distance calculation in this study. X_1 and X_2 represented the coordinates of points. The use of the absolute value (| . |) ensured that the result was always positive, in line with the interpretation of distance. Based on this calculation, a formula was derived to compute the manipulated features, as shown in Table 6.

3.4. Model Evaluation.

To assess the performance of this model in its regression task, the researcher employed Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The primary objective for both metrics was to achieve minimal values, indicating high precision in predictions. The novelty of this research lay in the application of RNN models and their variants to train complex engineered feature data, such as eye openness, upper and lower lip distances, lip length and height, and so forth. The performance results of this model were compared with metrics from baseline studies [1], [15], which implemented XGBoost and LSTM models for multimodal features.

$$\text{RMSE} = \sqrt{\frac{\sum(y_i - y_p)^2}{n}} \tag{2}$$

$$\text{MAE} = \frac{\sum|y_i - y_p|}{n} \tag{3}$$

$$y_i = \text{actual value}$$

$$y_p = \text{predicted value}$$

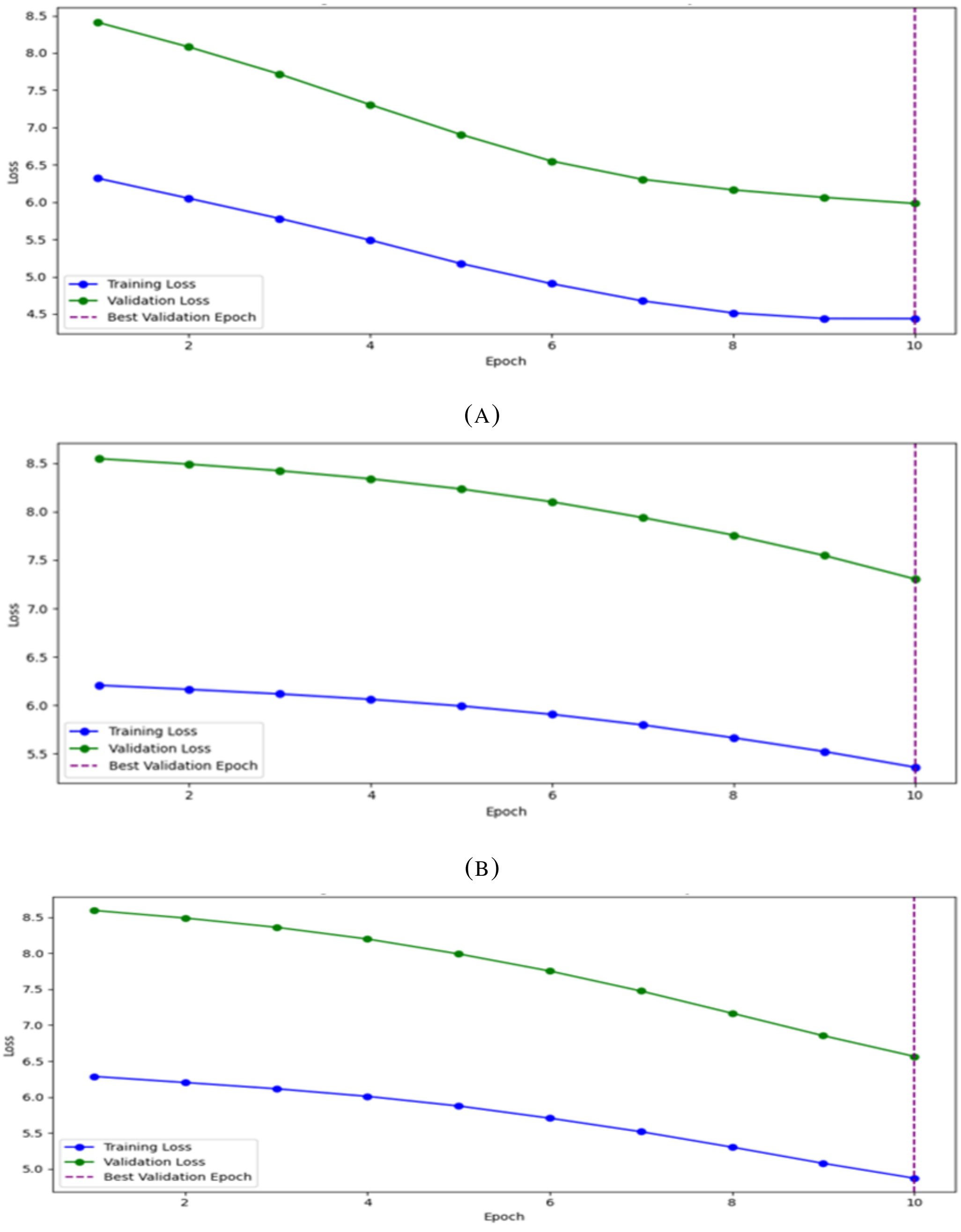
$$n = \text{number of observations}$$

4. Results and Discussion

In this section, the researcher presents the model learning outcomes using the first derivative and feature engineering data in illustrations Figure 5 and Figure 6. The performance is compared using the calculation of MAE and RMSE metrics. Additionally, in this section, the researcher discussed the comparison between the baseline study, and this proposed method.

In Figure 5, the performance graph of the three models on the first derivative data continues to show a decline, indicating the potential to achieve lower MAE and RMSE values. The RNN model achieves the best performance with 5.20 MAE and 6.07 RMSE, the lowest values among the models used for the first derivative. However, the GRU model has higher MAE and RMSE than LSTM. Furthermore, on the graph, the GRU model demonstrated better performance compared to LSTM. This can be observed through a smaller difference between training loss and validation loss. The performance of this model is influenced by its simplicity, indicating a need for increased complexity to achieve optimal results.

Figure 6 shows that the feature engineering model outperforms the first derivative model. In this case, the difference between training loss and validation loss is not as pronounced as in the plot of Figure 5, and the graph continues to show a decreasing trend, suggesting that the MAE and RMSE values could be minimized. The RNN model has the lowest MAE and RMSE values, at 5.04 and 6.03, respectively. The LSTM model may have a lower MAE but a higher RMSE than the GRU model.

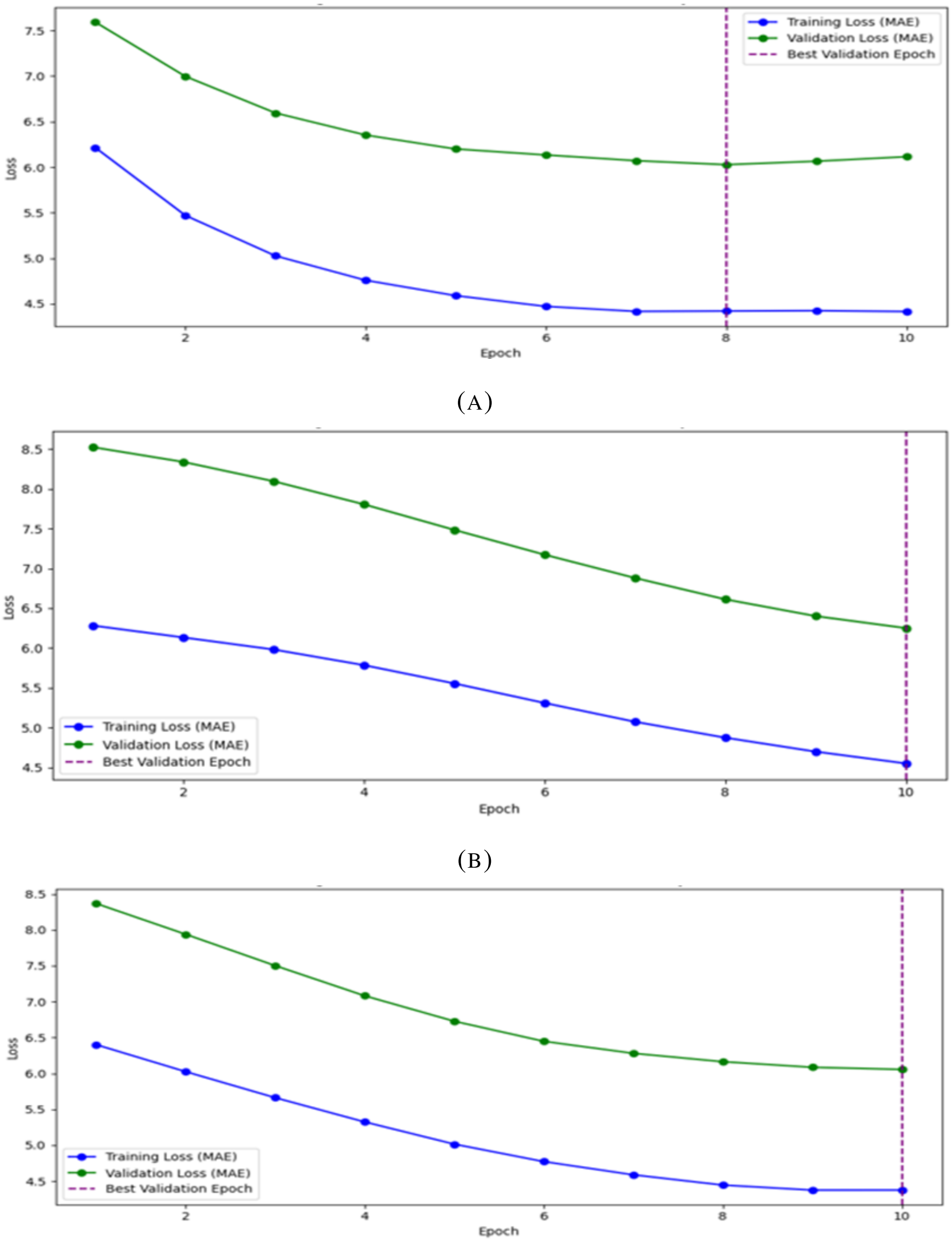


(A)

(B)

(C)

FIGURE 5. First derivative training results: (A) RNN, (B) LSTM, (C) GRU



(A)

(B)

(C)

FIGURE 6. Feature engineering training results: (A) RNN, (B) LSTM, (C) GRU

However, it is important to note that the model training in this study was limited to only 10 epochs due to hardware constraints. This limitation implies that the models might not have fully converged yet, as evidenced by the continuous decrease in both training and validation losses in the plotted graphs. Ideally, a higher number of epochs would allow the models to learn more deeply and potentially improve performance further. The evaluation results are presented in Table 7.

TABLE 7. Comparison of our best model with previous studies

Method	MAE	RMSE
XGBoost (baseline) [1]	5.28	6.22
LSTM (baseline) [15]	4.83	5.76
RNN + first derivative (ours)	5.20	6.07
LSTM + first derivative (ours)	5.72	7.89
GRU + first derivative (ours)	5.27	6.98
RNN + feature engineering (ours)	5.04	6.03
LSTM + feature engineering (ours)	5.08	6.42
GRU + feature engineering (ours)	5.09	6.02

In Table 7, there is a performance comparison between the proposed method and the baseline method. In Rumahorbo et al. [1], a first derivative method with a machine learning model was proposed. However, the data patients were not segmented uniformly, unlike in the proposed study, where segments were divided into 5-minute intervals. This inconsistency in training data hindered the model from learning optimally compared to uniformly segmented data. In Rasipuram et al. [15], multimodality was used, combining video, audio, and text in model training. However, the video modality itself still needs further development. Thus, if one modality reaches its maximum potential, it does not rule out the possibility that multimodality may yield even better results. Therefore, in the proposed study, the focus is on video and using feature engineering techniques to potentially enhance the results. Based on the researcher's findings, the eye and lip areas are relevant in detecting depression. Hence, the focus is on these areas, and feature manipulation is conducted to process them in learning.

Based on the conducted research, the researcher found that the eye and lip areas, as well as manipulated features, correlate with the severity of an individual's depression. The feature engineering method tends to offer new insights for future research, as it can provide quite optimal results. Additionally, based on the performance observed in Table 7, the proposed method, particularly feature engineering, excels in video data analysis, as it exhibits minimal MAE and RMSE values. Lower metric values indicate fewer errors in model learning. In Rasipuram et al. [14], it can provide insights and new developments regarding the utilization of multimodality using feature engineering. When each modality's potential, including video, audio, and text data, is harnessed effectively, it leads to optimal outcomes in analyzing or detecting multimodalities.

In Figure 5 and Figure 6, there is potential for further development of the model's learning. However, the experiments encountered challenges due to resource limitations, which necessitated simplifying the model's complexity. Therefore, researchers may consider increasing the complexity of the model by adding more layers, neurons, and epochs. Overall, the plotted graphs presented have not converged yet, indicating potential for improvement in terms of training and validation loss. To address this, an increase in the number of epochs is necessary. Additionally, the current RNN model outperforms other proposed methods in terms of performance. This may be because the model is simple, which prevents RNN from experiencing vanishing gradients. However, further experiments are needed to enhance the model's complexity for optimal results.

This study has provided insights for future research development. Assessing the severity of depression based on human facial expressions is a complex task. In this regard, the researcher has discovered that calculations involving lip width and height, the distance between eyelids when opening and closing, as well as the distance between the upper and lower lips, correlate with the severity of depression. Future research could further explore the areas of the eyes and lips, enriching the features used to train the model.

In conclusion, the proposed method, especially feature engineering, offers new insights into identifying the severity of depression. This method outperforms previous studies on DAIC-WOZ video data. The selection of features is a crucial aspect of the research. Based on the conducted study, the areas of the eyes and lips emerge as relevant for utilization in this field, thus warranting further exploration to enrich and enhance model learning with optimal outcomes.

5. Conclusion

This research conducted experiments to detect the severity of depression in individuals using first derivative and feature engineering. The findings provided insights that the areas around the eyes and lips have a significant correlation with an individual's level of depression, prompting researchers to focus more on these areas by applying feature engineering. This involved calculations such as measuring the distance between the upper and lower lips, lip length, lip height, and eye openness for both the right and left eyes. Based on the obtained MAE and RMSE results, feature engineering techniques, particularly with the RNN model, demonstrated superior performance with values of 5.04 MAE and 6.03 RMSE compared to the first-order derivative. This research outperformed other unimodal study, Rumahorbo et al. [1], indicating that feature engineering methods are quite effective in detecting depression based on facial expressions. However, when compared to multimodal approaches Rasipuram et al. [15], this research method showed higher MAE and RMSE values. Nevertheless, this research can provide insights to future researchers for processing video data using feature engineering techniques, aiming for optimal results from its multimodality. However, the challenge in this research lies in the model's simplicity due to resource limitations. Therefore, further experiments are needed to increase the model's complexity, such as adding layers and neurons, with the hope of improving overall performance. Additionally, there is still much to explore in terms of feature engineering, as detecting depression in humans is complex and influenced by various factors.

Data Availability

This study used the publicly available DAIC-WOZ dataset.

Author Contributions

Brilyan Nathanael Rumahorbo: Conceptualization, Methodology, Investigation, Writing – Original Draft

Gregorius Natanael Elwirehardja: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Project Administration

Bens Pardamean: Resources, Writing - Review & Editing, Supervision

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] B.N. Rumahorbo, K. Nanggala, G.N. Elwirehardja, B. Pardamean, Analyzing Important Statistical Features from Facial Behavior in Human Depression Using XGBoost, Commun. Math. Biol. Neurosci., 2023 (2023), 35. https://doi.org/10.28919/cmbn/7916.

[2] W. Marx, B.W.J.H. Penninx, M. Solmi, T.A. Furukawa, J. Firth, et al., Major Depressive Disorder, Nat. Rev. Dis. Prim. 9 (2023), 44. https://doi.org/10.1038/s41572-023-00454-1.

[3] M. Yuan, B. Yang, G. Rothschild, J.J. Mann, L.D. Sanford, et al., Epigenetic Regulation in Major Depression and Other Stress-Related Disorders: Molecular Mechanisms, Clinical Relevance and Therapeutic Potential, Signal Transduct. Target. Ther. 8 (2023), 309. https://doi.org/10.1038/s41392-023-01519-z.

[4] Y. Jiang, Y. Lu, Y. Cai, C. Liu, X. Zhang, Prevalence of Suicide Attempts and Correlates among First-Episode and Untreated Major Depressive Disorder Patients with Comorbid Dyslipidemia of Different Ages of Onset in a Chinese Han Population: A Large Cross-Sectional Study, BMC Psychiatry 23 (2023), 10. https://doi.org/10.1186/s12888-022-04511-z.

[5] M. Jacobs, M.F. Pradier, T.H. McCoy, R.H. Perlis, F. Doshi-Velez, et al., How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection, Transl. Psychiatry 11 (2021), 108. https://doi.org/10.1038/s41398-021-01224-x.

[6] W. Yan, Q. Ruan, K. Jiang, Challenges for Artificial Intelligence in Recognizing Mental Disorders, Diagnostics 13 (2022), 2. https://doi.org/10.3390/diagnostics13010002.

[7] G. Rughani, T.I.L. Nilsen, K. Wood, F.S. Mair, J. Hartvigsen, et al., The ¡scp¿selfBACK¡/scp¿ Artificial Intelligence-based Smartphone App Can Improve Low Back Pain Outcome Even in Patients with High Levels of Depression or Stress, Eur. J. Pain 27 (2023), 568–579. https://doi.org/10.1002/ejp.2080.

[8] S. Choudhary, M.K. Bajpai, K.K. Bharti, Computer Aided Diagnostic System with Reduced Electrode Set for Depression Detection Using Spatio-Temporal Attention Mechanism, Measurement 239 (2025), 115500. https://doi.org/10.1016/j.measurement.2024.115500.

[9] Q. Wang, L. Li, L. Qiao, M. Liu, Adaptive Multimodal Neuroimage Integration for Major Depression Disorder Detection, Front. Neuroinformatics 16 (2022), 856175. https://doi.org/10.3389/fninf.2022.856175.

[10] K.D. Kannan, S.K. Jagatheesaperumal, R.N.V.P.S. Kandala, M. Lotfaliany, R. Alizadehsanid, et al., Advancements in Machine Learning and Deep Learning for Early Detection and Management of Mental Health Disorder, arXiv:2412.06147 (2024). https://doi.org/10.48550/arXiv.2412.06147.

[11] N.K. Iyortsuun, S. Kim, M. Jhon, H. Yang, S. Pant, A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis, Healthcare 11 (2023), 285. https://doi.org/10.3390/healthcare11030285.

[12] B. Pardamean, H. Soeparno, A. Budiarto, B. Mahesworo, J. Baurley, Quantified Self-Using Consumer Wearable Device: Predicting Physical and Mental Health, Healthc. Inform. Res. 26 (2020), 83–92. https://doi.org/10.4258/hir.2020.26.2.83.

[13] I.D. Mienye, T.G. Swart, G. Obaido, Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications, Information 15 (2024), 517. https://doi.org/10.3390/info15090517.

[14] E.P. Wahyuddin, R.E. Caraka, R. Kurniawan, W. Caesarendra, P.U. Gio, et al., Improved LSTM Hyperparameters Alongside Sentiment Walk-Forward Validation for Time Series Prediction, J. Open Innov.: Technol. Mark. Complex. 11 (2025), 100458. https://doi.org/10.1016/j.joitmc.2024.100458.

[15] S. Rasipuram, J.H. Bhat, A. Maitra, B. Shaw, S. Saha, Multimodal Depression Detection Using Task-Oriented Transformer-Based Embedding, in: 2022 IEEE Symposium on Computers and Communications (ISCC), IEEE, 2022, pp. 01-04. https://doi.org/10.1109/ISCC55528.2022.9913044.

[16] R.P. Thati, A.S. Dhadwal, P. Kumar, S. P, A Novel Multi-Modal Depression Detection Approach Based on Mobile Crowd Sensing and Task-Based Mechanisms, Multimed. Tools Appl. 82 (2022), 4787–4820. https://doi.org/10.1007/s11042-022-12315-2.

[17] B.N. Rumahorbo, B. Pardamean, G.N. Elwirehardja, Exploring Recurrent Neural Network Models for Depression Detection Through Facial Expressions: A Systematic Literature Review, in: 2023 6th International Conference of Computer and Informatics Engineering (IC2IE), IEEE, 2023, pp. 209-214. https://doi.org/10.1109/IC2IE60547.2023.10331094.

[18] S.A.S. Mola, T.D.I.D. Ole, A.S. Karnyoto, et al. Fine-Tuning VGG16 Model for Driver Behavior Classification, Commun. Math. Biol. Neurosci. 2025 (2025), 47. https://doi.org/10.28919/cmbn/9165.

[19] M. Muzammel, H. Salam, A. Othmani, End-To-End Multimodal Clinical Depression Recognition Using Deep Neural Networks: A Comparative Analysis, Comput. Methods Programs Biomed. 211 (2021), 106433. https://doi.org/10.1016/j.cmpb.2021.106433.

[20] R. Flores, M.L. Tlachac, E. Toto, E. Rundensteiner, AudiFace: Multimodal Deep Learning for Depression Screening, in: Proceedings of the 7th Machine Learning for Healthcare Conference, PMLR 182, pp. 609–630, 2022.

[21] F. Ceccarelli, M. Mahmoud, Multimodal Temporal Machine Learning for Bipolar Disorder and Depression Recognition, Pattern Anal. Appl. 25 (2021), 493–504. https://doi.org/10.1007/s10044-021-01001-y.

[22] Y. Cao, Y. Hao, B. Li, J. Xue, Depression Prediction Based on Biattention-Gru, J. Ambient. Intell. Humaniz. Comput. 13 (2022), 5269–5277. https://doi.org/10.1007/s12652-021-03497-y.

[23] J. Gratch, R. Artstein, G. Lucas, et al. The Distress Analysis Interview Corpus of Human and Computer Interviews, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association, pp. 3123–3128, 2014.

[24] A. Saidi, S.B. Othman, S.B. Saoud, Hybrid Cnn-Svm Classifier for Efficient Depression Detection System, in: 2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), IEEE, 2020, pp. 229-234. https://doi.org/10.1109/IC_ASET49463.2020.9318302.

[25] U. Arioz, U. Smrke, N. Plohl, I. Mlakar, Scoping Review on the Multimodal Classification of Depression and Experimental Study on Existing Multimodal Models, Diagnostics 12 (2022), 2683. https://doi.org/10.3390/diagnostics12112683.

[26] F. Fischer, D. Zocholl, G. Rauch, B. Levis, A. Benedetti, et al., Prevalence Estimates of Major Depressive Disorder in 27 European Countries from the European Health Interview Survey: Accounting for Imperfect Diagnostic Accuracy of the PHQ-8, BMJ Ment. Health 26 (2023), e300675. https://doi.org/10.1136/bmjment-2023-300675.

[27] S. Nahulanthran, L. Tian, D. Kulić, M. Vered, Explaining Facial Expression Recognition, arXiv:2501.15864 (2025). https://doi.org/10.48550/arXiv.2501.15864.