IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 2, April 2025, pp. 1302~1310
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i2.pp1302-1310

Depression detection through transformers-based emotion recognition in multivariate time series facial data

Kenjovan Nanggala[1], Gregorius Natanael Elwirehardja[2,3], Bens Pardamean[1,2]
[1]Master of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia
[2]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta, Indonesia
[3]Department of Computer Science, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia

Article Info

Article history:

Received Nov 12, 2023
Revised Nov 16, 2024
Accepted Nov 24, 2024

Keywords:

Deep learning
Facial behavior analysis
Facial data
Major depressive disorder
Transformers

ABSTRACT

Globally, the prevalence of mental health disorders, particularly depression, has become a pressing issue. Early detection and intervention are vital to mitigate the profound impact of depression on individuals and society. Leveraging transformer models, renowned for their excellence in natural language processing and time series tasks, we explore their application in depression detection using multivariate time series (MTS) data from facial expressions. Transformer models excel in sequential data processing but remain relatively unexplored in facial expression analysis. This study aims to compare transformer models applied to first-order time derivative data with traditional methods. We use the distress analysis interview corpus wizard of oz (DAIC-WOZ) dataset and evaluate models with mean absolute error (MAE) and root mean squared error (RMSE) metrics. Results show that transformer models on first derivatives outperform others with an MAE of 4.42 and RMSE of 5.42. While transformer models on raw data surpass XGBoost in RMSE, they fall short of LSTM+transformer with an MAE of 5.41 and RMSE of 6.02. Preprocessing through differentiation enhances transformer models' ability to capture temporal patterns, promising improved depression detection accuracy.

Corresponding Author:

Kenjovan Nanggala
Master of Computer Science, School of Computer Science, Bina Nusantara University
Jakarta 11480, Indonesia
Email: kenjovan.nanggala@binus.ac.id

1. INTRODUCTION

Mental health disorders, particularly depression, have evolved into a significant global concern, impacting millions of individuals worldwide in recent years [1], [2]. Even if it is already quite severe, it can lead to cases of suicide [3]. The critical role of early detection and intervention in alleviating the profound consequences of depression on individuals and society at large is widely acknowledged [4], [5]. In the era of artificial intelligence and deep learning, researchers are actively exploring innovative approaches to enhance the accuracy and efficiency of depression detection [6].

This study aspires to contribute to the expanding body of knowledge in the field of mental health assessment, with a specific focus on the detection of depression. To accomplish this, we harness transformer models, a deep neural network architecture renowned for its exceptional performance in a variety of natural language processing (NLP) and time series data tasks [7], [8]. Transformers model employ attention mechanisms and self-attention layers to grasp the connections between words and phrases in a provided text [9]. While transformer models have garnered considerable attention for their effectiveness in processing sequential data, their application in the domain of facial expression analysis, particularly in the context of multivariate time series (MTS) data, has remained relatively uncharted territory [10].

This research is grounded in the following central research question: why do the regression outcomes resulting from the utilization of a transformer model on first-order time derivative data of facial behavior in depression detection compare to those of the basic method? The research endeavor encompasses an exploration of the potential of transformer models in the realm of depression detection through the analysis of MTS data derived from facial expressions [11]. The choice of utilizing a transformer model for this research stems from its ability to effectively capture complex sequential patterns and dependencies in MTS data, making it a promising approach for the task of depression detection. Furthermore, our aim is to evaluate the performance of transformer models in comparison to baseline models, with a particular emphasis on their ability to capture long-term temporal dependencies [12]. Additionally, this study seeks to assess the success of the transformer-based approach to depression detection as a valuable tool for early intervention and mental health support.

2. RELATED WORKS

Several notable studies in the domain of depression detection through facial analysis have paved the way for innovative approaches. Gavrilescu and Vizireanu [13] utilized the facial action coding system (FACS) to discern depression, anxiety, and stress (DASS) levels, achieving impressive accuracies of 87.2% for depression, 77.9% for anxiety, and 90.2% for stress with a unique three-layer architecture. Muzammel *et al.* [14] delved into major depressive disorder (MDD) detection, employing long short-term memory (LSTM) and convolutional neural network (CNN), with slightly improved accuracy of 66.25% compared to 65.60% for binary cases of depression. Grimm *et al.* [15] introduced the patient health questionnaire (PHQ-V) and generalized anxiety disorder (GAD-V), replacing PHQ-9 and GAD-7, employing three transformer blocks, leading to improved results. Sun *et al.* [16] crafted the deep feature fusion network (DFFN), achieving superior results in depression detection from a fusion of text, audio, and video modalities with a precision score of 0.91. Sun *et al.* [17] harnessed the transformer network to detect MDD, surpassing baseline methods with a concordance correlation coefficient (CCC) score of 0.733.

Rasipuram *et al.* [18] combined CNN and bidirectional long short-term memory (BiLSTM) layers for MDD detection, excelling in addressing imbalanced data, gender bias, and small-scale dataset challenges. Tigga and Garg [19] explored electroencephalogram (EEG) signals with an attention-based gated recurrent units transformer (AttGRUT) time series neural network, consistently outperforming other time-series models. Tiwary *et al.* [20] achieved the highest accuracy of 82% for MDD detection using a deep CNN model coupled with the DASS and FACS. Shangguan *et al.* [21] achieved an accuracy of 74.7% and a recall of 74.5% with the attention-based deep domain matching instance learning (ADDMIL) model for MDD detection. Guo *et al.* [22] combined the temporal dilated convolution network (TDCN) branches, excelling in terms of accuracy, recall, and F1-score, ranking second best for precision in automatic depression detection. These studies collectively contribute to the pursuit of accurate depression detection methods, signaling promising potential for the field.

The literature review provided valuable insights for your regression-based research on depression detection within MTS data using a transformer model. The implementation of the transformer model, inspired by studies like Sun *et al.* [17], remains a promising approach for modeling and regressing depression severity levels based on video-based MTS. Moreover, the experiences of handling data imbalance and bias, as addressed in [18], [23], can be invaluable in ensuring the robustness of your model when working exclusively with video data. Furthermore, an avenue for enhancing the model's performance lies in its adaptation and fine-tuning for video data, with the objective of minimizing mean absolute error (MAE) and root mean squared error (RMSE), as inspired by notable achievements in the field of depression detection as documented in the literature.

3. METHOD

In this study, we employ transformer models to detect depression from video data, with a particular focus on MTS facial behavior data. The choice of transformers is motivated by their proven effectiveness in handling MTS data, a characteristic not commonly explored in the context of facial expression analysis for depression detection. Compared to recurrent neural networks (RNNs) [24], transformers offer advantages in capturing long-range temporal dependencies [25], making them suitable for analyzing facial behavior over extended durations. This study aligns with previous research by Rasipuram *et al.* [18] that successfully utilized transformers in mental health disorder detection, primarily in audio and text modalities. Still, it aims to extend their application to the underexplored realm of time-series facial expression data. By capitalizing on the parallel processing capabilities of transformers, this study seeks to enhance the efficiency and accuracy of depression detection methods. The enhanced efficiency primarily relates to the model's ability to process data in parallel, which can lead to faster inference times and potentially reduce computational resources required for depression detection.

In Figure 1, it is illustrated that this paper consists of several sequential steps. These include data acquisition, which involves utilizing the distress analysis interview corpus wizard of oz (DAIC-WOZ) dataset, followed by preprocessing the obtained data. Subsequently, the process encompasses training and parameter tuning, model evaluation, and ultimately culminates in publishing the research findings.

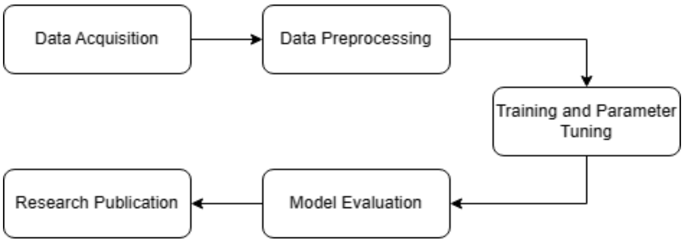
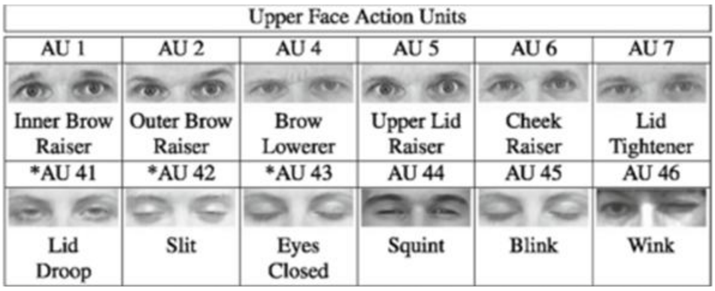


Figure 1. Research flow

3.1. Dataset

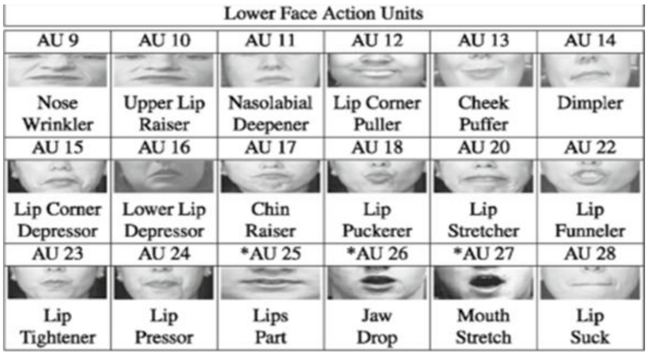
The dataset employed for this research is sourced from the DAIC-WOZ. This corpus consists of clinical interviews conducted by a virtual animated interviewer, designed for diagnosing psychological stress disorders, including depression. The dataset provides a rich source of real-world data for analyzing and understanding various aspects of these disorders. Utilizing such data facilitates comprehensive research and insights into the complexities of psychological distress.

The video data in this dataset comprises information related to facial action units that can be see in Figures 2 and 3, for the upper and lower face, respectively. The research flow encompasses data acquisition, preprocessing, feature selection using Pearson correlation (PC), model training and parameter tuning, and finally, model evaluation. The primary evaluation metrics include MAE and RMSE to assess the model's performance in predicting depression levels based on facial behavior data.



Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lid Tightener
*AU 41	*AU 42	*AU 43	AU 44	AU 45	AU 46
Lid Droop	Slit	Eyes Closed	Squint	Blink	Wink

Figure 2. Upper face facial action unit [26]



Lower Face Action Units					
AU 9	AU 10	AU 11	AU 12	AU 13	AU 14
Nose Wrinkler	Upper Lip Raiser	Nasolabial Deepener	Lip Corner Puller	Cheek Puffer	Dimpler
AU 15	AU 16	AU 17	AU 18	AU 20	AU 22
Lip Corner Depressor	Lower Lip Depressor	Chin Raiser	Lip Puckerer	Lip Stretcher	Lip Funneler
AU 23	AU 24	*AU 25	*AU 26	*AU 27	AU 28
Lip Tightener	Lip Pressor	Lips Part	Jaw Drop	Mouth Stretch	Lip Suck

Figure 3. Lower face facial action unit [26]

3.2. Data preprocessing

Before delving into the preprocessing steps, it's crucial to understand the pivotal role they play in refining raw data for analysis. Preprocessing encompasses various techniques aimed at enhancing the quality and reliability of the dataset, ensuring it is suitable for further analysis. These steps typically involve downloading the dataset, aggregation, cropping, splitting and mean calculation that can be seen in Figure 4. By executing preprocessing diligently, researchers can mitigate potential biases, address missing values, and standardize data formats, thereby laying a solid foundation for subsequent analyses. Hence, a comprehensive understanding of preprocessing techniques is indispensable for extracting meaningful insights from the data and making informed decisions based on analytical outcomes.

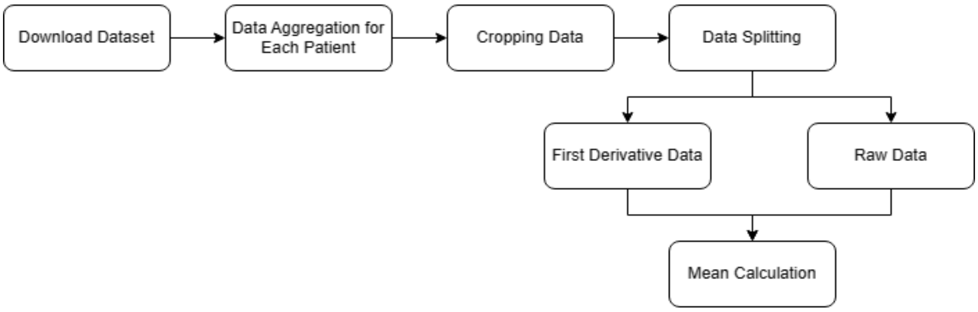


Figure 4. Data preprocessing flow

Refer to Figure 4, we can see that for the preprocessing stage, after consolidating text files from each participant into a dataframe with video recording coordinates, the researcher proceeds to crop the timestamp (start_time, stop_time) based on the audio recording data in the transcript file. The purpose is to align the timestamps in the combined dataframe with those in the audio recording to eliminate irrelevant or uninformative timestamps. In this study, all dataset features are utilized except for histogram of oriented gradients (HOG) features. The resulting cropped data is then exported to comma-separated values (CSV). Furthermore, a first derivative of the cropped data is computed, representing the difference between consecutive frames, and these datasets are also exported to CSV. The first derivative aims to investigate whether the rate of change in landmark positions or specific features significantly differs between normal individuals and those with depression. Afterward, the mean values for each dataset are calculated for feature selection, and only these mean values are used for the selection process. Subsequently, the selected features are used in the MTS data for modeling. Following data preprocessing, feature selection is performed using PC based on features with significant p-values, considering both the raw and first derivative (mean) data. Utilizing PC for feature selection helps identify features with strong correlations, reducing dataset dimensions, enhancing computational efficiency, and improving model performance by retaining the most informative and relevant features in data analysis.

Following the described data preprocessing steps, feature selection was performed using PC, considering features with p-values <0.1 for both the raw and first derivative (mean) data. This method helps identify features exhibiting strong correlations, leading to a reduction in dataset dimensions. Consequently, it enhances computational efficiency and model performance by retaining the most informative and relevant features for further analysis. Subsequently, the results of the feature selection, including the p-value of each feature, are presented following.

The data presented in Table 1 highlights the outcomes of feature selection based on a criterion of p-values below 0.1, albeit still approaching 0.05. Notably, AU12_r exhibits a p-value of 0.012, representing a significant feature associated with the lip area. Additionally, y_h0 and y_h1, with p-values of 0.069 and 0.061 respectively, are identified as crucial components of eye gaze within the dataset.

Table 1. Results of feature selection

Features	P-Value
AU12_r	0.012
y_h0	0.069
y_h1	0.061

3.2. Model architecture

The model's architecture serves as the blueprint that outlines its structure, including the arrangement of layers, connections, and algorithms employed for processing data. Understanding the architecture provides insights into how the model learns from input data, makes predictions, and adapts to varying complexities. Additionally, it sheds light on the model's capabilities, limitations, and potential for optimization. Hence, exploring the architecture is pivotal for comprehending the model's inner workings and its efficacy in addressing the research objectives. In this research, we use transformer as the model and the architecture of the model can be seen in Figure 5.

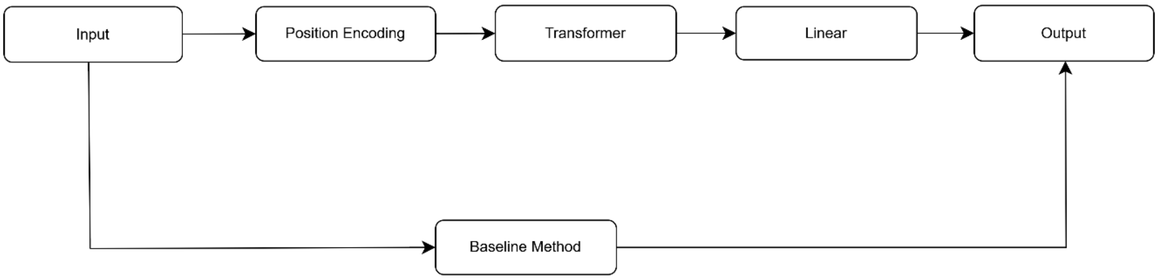


Figure 5. Transformer model architecture

From Figure 5, we can see that this study utilizes a transformer model architecture that includes position encoding, self-attention layers, feed-forward neural network layers, and linear activation functions for depression level predictions. Hyperparameter tuning, performed through grid search, allows for optimal model configuration. The research leverages insights from previous work in the domain of depression detection while extending the application of transformers to improve the accuracy and effectiveness of the analysis of MTS facial behavior data, ultimately contributing to the development of more reliable and efficient depression detection methods. The following hyperparameters were explored during the grid search process:

- Learning rate: [0.05, 0.01, 0.001]
- Neurons in each layer: [32, 64, 128, 256]
- Number of attention heads: [1, 2, 4, 8]
- Linear units: [512, 1024, 2048, 4096]
- Dimension of the feed-forward (DFF): [64, 128, 256]
- Dropout rate: [0.1, 0.2, 0.3, 0.4]

For the best parameter settings used in this study are as follows:

- Learning rate: 0.05
- Neurons in each layer: 128
- Number of attention heads: 2
- Linear units: 2048
- DFF: 64
- Dropout rate: 0.1

3.3. Model evaluation

In evaluating the performance of the proposed transformer model on first derivative video data, we employ two primary metrics: MAE and RMSE. These metrics are chosen for their ability to quantify the average magnitude of the errors in a set of predictions without considering their direction. These metrics are well-suited for our study as they offer a clear and straightforward interpretation of model performance. MAE provides a natural and easily interpretable measure of average error magnitude, while RMSE gives a higher weight to larger errors, which can be particularly relevant in scenarios where large errors are more detrimental than smaller ones.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

where: $\hat{y} = predicted\ value\ of\ y$, $\underline{y} = mean\ value\ of\ y$, and $n = number\ of observations/rows$.

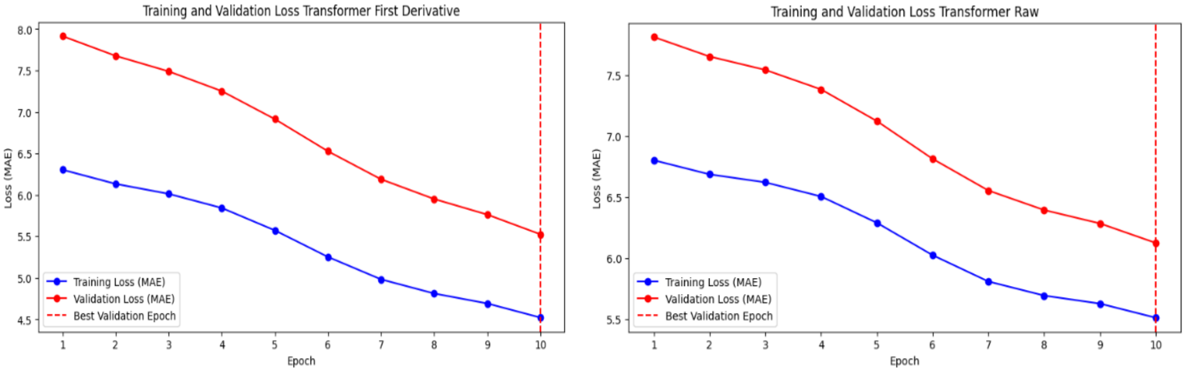
The novelty of our research lies in the implementation of transformer models to process first derivative data extracted from video sequences. This application is innovative, as transformers are predominantly used in NLP and have only recently been explored in the context of video data analysis. By applying the transformer's self-attention mechanisms to the temporal dynamics captured in the first derivative of video data, our model aims to enhance the learning of temporal patterns that are crucial for accurate predictions. By comparing our transformer model's performance on first derivative video data against these baselines using MAE and RMSE, we aim to demonstrate its efficacy and the potential advantages of our approach for video-based applications. This comparison allows us to establish the transformer model's robustness and accuracy, further contributing to the domain of video data analysis and expanding the utility of transformer architectures beyond their traditional domains.

4. RESULT AND DISCUSSION

In this section, we meticulously dissect the findings obtained from our research methodology and delve into their implications. Through a comprehensive examination of the results, we aim to unravel the underlying patterns, trends, and correlations embedded within the data. Furthermore, we engage in a critical discourse to contextualize our findings within the existing body of knowledge, offering insights into their significance and potential contributions to the field. By intertwining the presentation of results with in-depth discussions, we strive to provide a holistic understanding of the research outcomes and their broader implications.

4.1. Result

In this research, the researcher utilized a transformer model on first derivative data as well as raw data, which can be seen in Figure 6. Figure 6(a) training and validation loss transformer first derivative and Figure 6(b) training and validation loss transformer raw. The results show a consistent decline in both data for both training and validation MAE loss over the epochs, indicating the model's learning progress. However, it is important to note that the model has not yet converged to its optimal performance. While the decline in validation loss alongside training loss suggests promising generalization and a lack of overfitting, the transformer's performance on the first derivative data is still showing room for improvement in terms of MAE and RMSE values. The transformer model achieved notable results with an MAE of 4.42 and an RMSE of 5.42, but it hasn't reached its convergence point. It is worth mentioning that the transformer with the first derivative model outperforms the baseline models. On the other hand, the transformer model using raw data shows better results compared to the baseline extreme gradient boosting (XGBoost) model in terms of RMSE, but it is slightly behind the baseline LSTM+transformer model, with an MAE of 5.41 and an RMSE of 6.02. The evaluation results are presented in Table 2.



(a) (b)

Figure 6. Plotting results of the training using transformer model on data (a) first derivative and (b) raw

The results depicted in the plots for the transformer model applied to the first derivative and raw data demonstrate significant findings. The transformer model’s performance on the first derivative data is notably superior, achieving a lower MAE of 4.42, which surpasses the benchmark models. This lower MAE suggests strong predictive accuracy and could indicate that preprocessing data through differentiation may help the transformer model capture underlying trends more effectively. For the raw data, the transformer model still outperforms the XGBoost baseline in terms of RMSE but does not exceed the LSTM+transformer model's performance.

Table 2. Comparison of the proposed model with the baseline model

Parameter	MAE	RMSE	R^2
Transformer raw data (our model)	5.41	6.02	0.20
Transformer first derivative (our model)	4.42	5.42	0.22
XGBoost first derivative [23]	5.41	6.32	0.02
LSTM+transformer raw data [18]	4.83	5.76	-

4.2. Discussion

This finding could suggest that while the transformer model is robust, the incorporation of LSTM into the architecture might capture temporal dependencies in the raw data more effectively, which are perhaps less pronounced after the differentiation process. Such a result suggests that using first derivative data in this context might capture relevant information that allows the transformer model to predict more accurately compared to its baseline counterparts or even the raw data. It also validates the effectiveness of the transformer architecture in handling this type of time-series data, potentially providing insights that could be leveraged in other similar applications. However, it would be important to look at additional metrics and conduct further tests, such as cross-validation, to confirm these findings and ensure the model's robustness.

A deeper analysis of the first derivative data reveals that the model effectively captures patterns of change within the videos, as evidenced by the relatively low values of MAE and RMSE. However, the still relatively low R^2 values indicate that the model using first derivative data can explain slightly more variation in the target data compared to the model using raw data. This may be due to the first derivative process's role in reducing noise or fluctuations present in the raw video data. By eliminating or reducing small or fluctuating aspects that may not be relevant, the model can focus more on larger and more significant patterns of change. Additionally, the first derivative may be more effective in highlighting dynamic changes in data over time, providing an advantage in detecting small changes or trends that may be overlooked by models using raw data.

Meanwhile, when using raw data, the model struggles to capture patterns that may exist in video changes. This is reflected in the higher MAE and RMSE values, indicating that the model has a higher level of error in predicting the severity of depression levels from unprocessed video data. However, for R^2 metrics, both data types yield relatively low values. This may be due to limitations on the amount of data, as Transformers often require large amounts of data for effective training. If the dataset used is relatively small or not sufficiently representative of the variations that may occur in the population, then the model may not generalize well to new data.

5. CONCLUSION

In conclusion, our study explores the potential of transformer models in detecting signs of depression through the analysis of video data, with a specific focus on facial behavior as a representation of MTS. Experimental results demonstrate that the application of the transformer model to video data, particularly after processing with the first derivative, yields superior performance compared to raw data. The transformer model exhibits improved performance metrics, including MAE of 4.42, RMSE of 5.42, and R^2 value of 0.22, outperforming baseline models and underscoring its capability to capture relevant temporal patterns for enhanced predictive accuracy. While the study contributes significantly to understanding the potential of transformer models in depression detection through video data, addressing limitations such as computational requirements and dataset availability is essential. Opportunities for future research lie in expanding datasets to improve model generalization, integrating findings with mental health platforms or online counseling services, and exploring alternative approaches, such as feature engineering, to enhance the model's capabilities further. However, limitations of the study involve the transformation of video data, which may not fully encompass complex non-verbal aspects, and the need for larger datasets to enhance model generalization. To mitigate these limitations, efforts are required to obtain broader datasets for improved model generalization. Future research opportunities include expanding datasets, integrating research findings with mental health platforms or online counseling services, and conducting research without using feature selection to measure the impact of the process and maintain all available features in the dataset. Additionally, the development of this research can be enhanced through feature engineering, where specific and relevant features such as interocular distance, lip height, and lip width can provide more informative data representations. Feature engineering offers significant advantages over previous approaches, allowing for a deeper understanding of depression-related patterns and potentially improving the model's ability to detect depression signs.

REFERENCES

[1] C. Su, Z. Xu, J. Pathak, and F. Wang, “Deep learning in mental health outcome research: a scoping review,” *Translational Psychiatry*, vol. 10, no. 1, Apr. 2020, doi: 10.1038/s41398-020-0780-3.

[2] L. Baer and M. A. Blais, *Handbook of clinical rating scales and assessment in psychiatry and mental health*. Totowa, NJ: Humana Press, 2010, doi: 10.1007/978-1-59745-387-5.

[3] G. N. Elwirehardja, M. Isnan, A. S. Perbangsa, K. Muchtar, and B. Pardamean, “Trends, opportunities, and challenges in detecting depressive disorders through mobile devices: a review,” in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, Aug. 2023, pp. 188–193, doi: 10.1109/COSITE60233.2023.10249859.

[4] M. Briley and J.-P. Lépine, “The increasing burden of depression,” *Neuropsychiatric Disease and Treatment*, vol. 7, no. 1, May 2011, doi: 10.2147/NDT.S19617.

[5] A. Murru *et al.*, “The implications of hypersomnia in the context of major depression: results from a large, international, observational study,” *European Neuropsychopharmacology*, vol. 29, no. 4, pp. 471–481, Apr. 2019, doi: 10.1016/j.euroneuro.2019.02.011.

[6] A. M. Nezu, K. S. McClure, and C. M. Nezu, “The assessment of depression,” in *Treating Depression*, Wiley, pp. 24–51, 2015, doi: 10.1002/9781119114482.ch2.

[7] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, “Overview of the transformer-based models for NLP tasks,” *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 179-183, doi: 10.15439/2020F20..

[8] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, “Transformers in vision: a survey,” *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–41, Jan. 2022, doi: 10.1145/3505244.

[9] K. Nanggala, G. N. Elwirehardja, and B. Pardamean, “Systematic literature review of transformer model implementations in detecting depression,” in *2023 6th International Conference of Computer and Informatics Engineering (IC2IE)*, pp. 203–208, Sep. 2023, doi: 10.1109/IC2IE60547.2023.10331448.

[10] T. W. Cenggoro, B. Mahesworo, A. Budiarto, J. Baurley, T. Suparyanto, and B. Pardamean, “Features importance in classification models for colorectal cancer cases phenotype in Indonesia,” *Procedia Computer Science*, vol. 157, pp. 313–320, 2019, doi: 10.1016/j.procs.2019.08.172.

[11] L. Kong, J. Yu, D. Tang, Y. Song, and D. Han, “Multivariate time series anomaly detection with generative adversarial networks based on active distortion transformer,” *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9658–9668, May 2023, doi: 10.1109/JSEN.2023.3260563.

[12] F. Zeng, M. Chen, C. Qian, Y. Wang, Y. Zhou, and W. Tang, “Multivariate time series anomaly detection with adversarial transformer architecture in the internet of things,” *Future Generation Computer Systems*, vol. 144, pp. 244–255, Jul. 2023, doi: 10.1016/j.future.2023.02.015.

[13] M. Gavrilescu and N. Vizireanu, “Predicting depression, anxiety, and stress levels from videos using the facial action coding system,” *Sensors*, vol. 19, no. 17, Aug. 2019, doi: 10.3390/s19173693.

[14] M. Muzammel, H. Salam, and A. Othmani, “End-to-end multimodal clinical depression recognition using deep neural networks: a comparative analysis,” *Computer Methods and Programs in Biomedicine*, vol. 211, Nov. 2021, doi: 10.1016/j.cmpb.2021.106433.

[15] B. Grimm, B. Talbot, and L. Larsen, “PHQ-V/GAD-V: assessments to identify signals of depression and anxiety from patient video responses,” *Applied Sciences*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/app12189150.

[16] G. Sun, S. Zhao, B. Zou, and Y. An, “Multimodal depression detection using a deep feature fusion network,” in *Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)*, Dec. 2022, doi: 10.1117/12.2662620.

[17] H. Sun *et al.*, “Multi-modal adaptive fusion transformer network for the estimation of depression level,” *Sensors*, vol. 21, no. 14, Jul. 2021, doi: 10.3390/s21144764.

[18] S. Rasipuram, J. H. Bhat, A. Maitra, B. Shaw, and S. Saha, “Multimodal depression detection using task-oriented transformer-based embedding,” in *2022 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–4, Jun. 2022, doi: 10.1109/ISCC55528.2022.9913044.

[19] N. P. Tigga and S. Garg, “Efficacy of novel attention-based gated recurrent units transformer for depression detection using electroencephalogram signals,” *Health Information Science and Systems*, vol. 11, no. 1, 2023, doi: 10.1007/s13755-022-00205-8.

[20] G. Tiwary, S. Chauhan, and K. K. Goyal, “Video based deep CNN model for depression detection,” *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 10, pp. 59–64, Oct. 2022, doi: 10.17762/ijritcc.v10i10.5735.

[21] Z. Shangguan, Z. Liu, G. Li, Q. Chen, Z. Ding, and B. Hu, “Dual-stream multiple instance learning for depression detection with facial expression videos,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 554–563, 2023, doi: 10.1109/TNSRE.2022.3204757.

[22] Y. Guo, C. Zhu, S. Hao, and R. Hong, “Automatic depression detection via learning and fusing features from visual cues,” *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2806–2813, Oct. 2023, doi: 10.1109/TCSS.2022.3202316.

[23] B. N. Rumahorbo, K. Nanggala, G. N. Elwirehardja, and B. Pardamean, “Analyzing important statistical features from facial behavior in human depression using XGBoost,” *Communications in Mathematical Biology and Neuroscience*, vol. 2023, pp. 1-24, 2023, doi: 10.28919/cmbn/7916.

[24] P. Tang, Q. Zhang, and X. Zhang, “A recurrent neural network based generative adversarial network for long multivariate time series forecasting,” in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pp. 181–189, Jun. 2023, doi: 10.1145/3591106.3592306.

[25] K. Muchtar, F. Rahman, T. W. Cenggoro, A. Budiarto, and B. Pardamean, “An improved version of texture-based foreground segmentation: block-based adaptive segmenter,” *Procedia Computer Science*, vol. 135, pp. 579–586, 2018, doi: 10.1016/j.procs.2018.08.228.

[26] R. Zhi, M. Liu, and D. Zhang, “A comprehensive survey on automatic facial action unit analysis,” *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, May 2020, doi: 10.1007/s00371-019-01707-5.

BIOGRAPHIES OF AUTHORS

Kenjovan Nanggala is a Master of Computer Science student at Bina Nusantara University, Indonesia. He enrolled in 2019 and is actively participating in the university's master track program as part of his academic journey toward obtaining a master's degree. His research interests include artificial intelligence, machine learning, deep learning, and mental health. He can be contacted at email: kenjovan.nanggala@binus.ac.id.

Gregorius Natanael Elwirehardja is adjunct research from Bioinformatics & Data Science Research Center and NVIDIA-BINUS Artificial Intelligence Research & Development Center of Bina Nusantara University, and a certified instructor at NVIDIA deep learning institute. He completed both his bachelor and master level education in Bina Nusantara University. His research interests include applied machine learning in various fields including, but are not limited to, computer vision, mental health, and natural language processing. He can be contacted at email: gregorius.elwirehardja@binus.ac.id.

Bens Pardamean has over forty years of global experience in information technology, bioinformatics, and education. His professional experience includes being a practitioner, researcher, consultant, entrepreneur, and lecturer. He currently holds a dual appointment as Director of Bioinformatics & Data Science Research Center (BDSRC) | AI Research & Development Center (AIRDC), and Professor of Computer Science at Bina Nusantara (BINUS) University in Jakarta, Indonesia. He earned a doctoral degree in informatics research from University of Southern California (USC), as well as a master's degree in computer education and a bachelor's degree in computer science from California State University at Los Angeles (USA). He can be contacted at email: bpardamean@binus.edu.