Swin Transformer with Enhanced Dropout and Layer-wise Unfreezing for Facial Expression Recognition in Mental Health Detection

Mujiyanto Mujiyanto
Department of Magister of Informatics Engineering, University AMIKOM Yogyakarta, Indonesia
mujiyanto@amikom.ac.id (corresponding author)

Arief Setyanto
Department of Magister of Informatics Engineering, University AMIKOM Yogyakarta, Indonesia
arief_s@amikom.ac.id

Kusrini Kusrini
Department of Magister of Informatics Engineering, University AMIKOM Yogyakarta, Indonesia
kusrini@amikom.ac.id

Ema Utami
Department of Magister of Informatics Engineering, University AMIKOM Yogyakarta, Indonesia
ema.u@amikom.ac.id

Received: 30 September 2024 | Revised: 25 October 2024 | Accepted: 3 November 2024

 DOI: https://doi.org/10.48084/etasr.9139

ABSTRACT

This study presents an improved Facial Expression Recognition (FER) model using Swin transformers for enhanced performance in detecting mental health through facial emotion analysis. In addition, some techniques involving better dropout and layer-wise unfreezing were implemented to reduce model overfitting. This study evaluates the proposed models on benchmark datasets such as FER2013 and CK+ and real-time Genius HR data. Model A has no dropout layer, Model B has focal loss, and Model C has enhanced dropout and layer-wise unfreezing. Model C was the best among all proposed models, achieving test accuracies of 71.23% on FER2013 and 78.65% on CK+. Weighted cross-entropy loss and image augmentation were used to handle class imbalance. Based on Model C emotion predictions, a scoring mechanism was designed to analyze employees' mental health for the next 30 days. The higher the score, the higher the risk of mental health. This study demonstrates a practical version of the Swin transformer in FER models for detecting and early mental health intervention.

Keywords-swin transformer; facial expression recognition; mental health detection; overfitting mitigation

I. INTRODUCTION

Mental health appears to have joined the front ranks of contemporary workplace issues. Approximately 15% of personnel will be affected by mental illness once at any time, the most obvious of which include depression and anxiety. The economic losses through loss of productivity and absenteeism are enormous, reaching up to 12 billion lost workdays and almost 1 trillion USD lost economic output per year [1]. In this respect, various recent developments, such as the Facial Expression Recognition (FER) method, have been implemented for early identification and intervention regarding mental disorders [2]. Incorporating FER into mental health assessments can prevent burnout or depression. As early-stage intervention can prevent these issues, there is an ever-growing demand for novel ways to conduct mental health assessments.

These systems play a vital role in understanding human feelings through machines that are operational in several sectors of daily life, including human-computer interaction, security, diagnosis of mental health, and social robotics. FER can help systems interpret emotional cues more precisely during human-computer interaction for personalized user experience offers [3]. During the assessment of mental health, FER helps clinicians estimate signs of emotional distress or disorders [4]. In security, FER can help monitor criminal behavior based on altered facial expressions. It also helps predict mental health disorders through emotion analysis, providing excellent early intervention [5].

Despite such a growth in deep learning, FER still faces several problems due to subtlety and diversity in human facial expression and other influential factors such as lighting conditions, head pose, and identity bias [6]. These issues can lead to increased overfitting while training a deep neural network based on small or imbalanced datasets [4, 7, 8]. Overfitting causes models to perform well on training data but poorly on new, unseen data, reducing their effectiveness in real-world applications [9, 10]. Techniques such as data augmentation, class weighting, and regularization have been used to overcome these challenges and enhance model robustness and generalization [11, 12].

Traditional FER methods relied on handcrafted features and simple classifiers, which struggled with high variations within classes and were sensitive to environmental changes [13, 14]. Recent approaches use deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn discriminative features [15, 16]. CNNs excel at learning spatial features from images, while RNNs focus on the temporal dynamics of facial expressions [17]. Recently, transformer-based models such as Vision Transformers (ViT) [18] and Swin transformers have shown great promise in computer vision by capturing broader contexts and handling long-range dependencies [12, 19]. With their hierarchical structure and efficiency in modeling local and global features, Swin transformers play a decisive role in FER [19], but their usage is still exploratory [19-24].

In addition, transformer models are usually overfitted with many parameters in FER research due to the limited size of datasets [25]. Several methods have been proposed to reduce overfitting by leveraging data augmentation and class-weighted loss functions due to class imbalance issues [8, 11]. Data augmentation increases the diversity of training data, while class weighting acts directly on loss, considering underrepresented classes as more critical [26]. However, few works combined such approaches with transformer architectures in FER. This study extends current state-of-the-art FER models using the Swin transformer architecture and new overfitting avoidance strategies [27]. In addition, Swin transformers are used in a comprehensive data augmentation that includes class-weighted loss functions [24]. In the proposed framework, the hierarchical Swin transformer feature representation is used to model facial features, both local and global, which is crucial to recognizing subtle facial expressions [28, 29]. In addition, various data augmentation techniques are employed to increase data variation and prevent overfitting, such as random rotation, flipping, color jittering, and erasing.

In this study, class weights are pre-calculated and added to the loss function to deal with class imbalance and enhance the model robustness [8]. Early stopping, learning rate scheduling, and clipping gradients are also used to improve the model and avoid overfitting issues [30]. Consequently, this results in a much better generalization compared to the direct utilization of baseline models. Furthermore, this study discusses how this model could find practical applications in assessing mental health. Predicting mental health status through emotion recognition is timely and noninvasive toward early workplace intervention.

II. MATERIALS AND METHODS

A. Dataset Collection

The datasets used in this study include FER2013 [31], CK+, and the Genius HR dataset. FER2013 consists of 48×48 grayscale images divided into seven emotion classes: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. CK+ contains 920 images across eight emotion classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise, Neutral, and Contempt. This dataset is split into 736 training images, with 92 images, each for validation and testing. Although FER2013 provides a rather large sample, CK+ is relatively small, containing only 920 images. Therefore, generalization based on this dataset alone may be difficult [32]. In addition, although helpful and convenient for testing in real-world conditions, the Genius HR dataset contains 500 images taken in an office of just five employees with a small amount of diversity in critical demographics, which could introduce bias and limit its application to larger groups of people [33].

Data augmentation was used to make the datasets more diverse and balanced. As FER datasets are both limited and very difficult to obtain, data augmentation allows us to simulate various demographic representations within the Genius HR dataset [8, 11]. Data augmentation aims to enhance the model's generalization capabilities despite the constraints of the available datasets. Table I summarizes the distribution of each dataset.

TABLE I. OVERVIEW OF DATASETS

Dataset	Resolution	Training	Test	Total
FER2013	48×48	28,709	7,178	35,887
CK+	48×48	736	184	920
Genius HR	48×48	400	100	500

B. Preprocessing and Augmentation Techniques

The preprocessing for both FER2013 and CK+ datasets included resizing all images to 48×48 pixels and then converting them to grayscale when needed. All CK+ images were first normalized to grayscale. These images were then individually normalized such that the pixel values have a zero mean and a unit variance - a common preprocessing feature to stabilize the training process. The normalization process can be expressed mathematically as follows:

$$x' = \frac{x-\mu}{\sigma} \quad (1)$$

where x' is the normalized pixel value, x is the original pixel value, μ is the mean, and σ is the standard deviation of the pixel values in the dataset. This step will maintain the pixel values within an appropriate scale and help to achieve faster model convergence during training by reducing the possible vanishing or exploding gradient problems [34]. The following data augmentation techniques were used on the datasets to increase robustness and avoid overfitting.

- Random horizontal flip introduces variability by flipping images horizontally, simulating mirrored facial expressions.
- Random rotation within a small degree range up to 15° introduces robustness against variations in head pose.

- Random erasing removes small sections of an image to simulate occlusions, improving generalization when handling masked or obscured faces.

These methods give a more diverse dataset and improve model generalizability.

Fig. 1. Augmented images from FER2013.

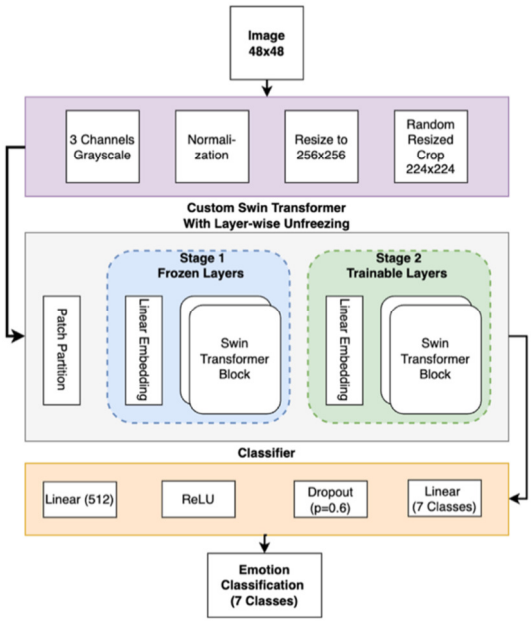

Fig. 2. Architecture of the proposed FER model.

C. Loss Functions and Optimization

The weighted cross-entropy loss was applied to handle class imbalance in the FER2013 and CK+ datasets. Class weights were calculated using:

$$w_c = \frac{N}{n_c} \quad (2)$$

where w_c is the weight for class c, N is the total number of samples, and n_c is the number of samples in class c. This method helps the model not to be biased towards the majority classes, giving better performance in all emotion categories. AdamW was used to decouple weight decay, effectively reducing overfitting and improving generalization [24], offering also stability and efficiency in model convergence.

D. Evaluation Metrics

The difference between training and validation accuracy/loss was monitored throughout the training process to assess overfitting:

$$RER = \frac{Validation\ Loss - Training\ Loss}{Training\ Loss} \quad (3)$$

A higher RER indicates that the model is overfitting by performing well on the training data but not generalizing effectively to the validation set. Key evaluation metrics include accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions, while precision quantifies the true positives out of predicted positives. Recall focuses on retrieving all relevant instances, and F1-score balances precision and recall.

E. Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM was employed to visualize the feature maps of the Swin transformer model, highlighting the regions in an image that contributed most to the model's predictions. The key expression used is:

$$L^c_{Grad\text{-}CAM} = ReLU(\sum_k \alpha_k^c A^k) \quad (4)$$

This technique utilizes the gradients flowing into the final convolutional layer to produce a localization map that shows the areas of the input image that influenced the classification. In the context of facial expression recognition, Grad-CAM allowed us to verify whether the model was focusing on relevant facial features, such as eyes or mouth, when making predictions.

F. Mental Health Scoring System

This study introduces a conceptual framework integrating FER models, trained on FER2013 and CK+ datasets, with real-world data from the Genius HR dataset. The primary goal of this system is to predict employee emotions and map them to a robust mental health scoring system, which allows for early detection of mental health risks. The model minimizes overfitting and optimizes generalization, ensuring that it can accurately predict emotions from facial images captured during daily attendance.

The core of this system involves calculating a Mental health Score (MS) for each employee based on the probability distribution of predicted emotions over a 30-day period. The daily average emotion prediction is recorded for each employee, and the mental health score is derived based on these emotion probabilities. The mental health score, denoted as S_i for an individual i, is calculated as a weighted sum of emotion probabilities:

$$S_i = \sum_{j=1}^{n} w_j \cdot P_{ij} \quad (5)$$

where P_{ij} represents the probability prediction of emotion j for individual i and w_j is the corresponding weight assigned to each emotion based on its impact on mental health. Negative emotions, such as anger and sadness, are assigned higher weights w_j due to their significant influence on well-being. In contrast, positive emotions are assigned lower weights, as they are generally associated with minimal risks to mental health.

Drawing from [35], the correlation between mental health and facial emotions is used to derive weights for each emotion. Simcock's work specifically relates to chronic depression (SP12), and Table II displays the correlation values used to inform the mental health score.

TABLE II. EMOTION-MENTAL HEALTH CORRELATION

Emotion	Correlation (SP12 depression)
Anger	-0.16
Fear	-0.06
Happy	0.15
Neutral	0.40
Sad	-0.04
Disgust	-0.03 (derived)
Surprise	-0.06 (derived)

The MS score is calculated using the following interpolation formula:

$$MS = 50 - 50 \cdot CR \quad (7)$$

where CR represents the correlation value for the detected emotion. This formula translates correlation values ranging from -1 to +1 into a mental health score between 0 and 100. A higher score indicates a greater mental health risk [5]. Table III outlines the mental health scores assigned to each emotion. In this scoring system, the Neutral emotion represents the lowest mental health risk, while Anger is associated with the highest risk. Over a 30-day period, the system aggregates daily emotional predictions for each individual to assess overall mental health trends.

TABLE III. MENTAL HEALTH SCORE DISTRIBUTION

Emotion	Mental Score (MS)
Disgust	53
Anger	58
Fear	53
Surprise	46
Happy	43
Sad	53
Neutral	30

The primary focus is to develop the most effective model that avoids overfitting by carefully selecting the optimal architecture and training strategies. Once the best model is identified, it will be used to predict emotional expressions from facial images in the Genius HR dataset on a routine basis during daily attendance.

III. RESULTS AND DISCUSSION

This section presents the experimental results comparing the three custom Swin transformer model architectures (Model A, Model B, and Model C) on the FER2013, CK+, and real-world Genius HR datasets. The discussion covers model performance in accuracy, precision, recall, and F1-score, along with confusion matrices.

A. Model Performance

Models A, B, and C were tested on FER2013 and CK+ datasets. Model C performed best among all the proposed models, with an overall accuracy of 72.5% for FER2013 and 95.3% for CK+. In contrast, Model A achieved an accuracy of 68.1% for FER2013 and 91.2% for CK+, and Model B achieved an accuracy of 69.7% for FER2013 and 93.5% for CK+, respectively. Table II shows all results.

TABLE IV. PERFORMANCE COMPARISON OF MODELS A, B, AND C ON FER2013 AND CK+ DATASETS

Model	Architecture	Dataset	Accuracy
A	No dropout	FER2013	70.05%
		CK+	95.51%
B	With Focal Loss	FER2013	64.61%
		CK+	83.15%
C	With enhanced dropout, layer-wise unfreezing	FER2013	71.23%
		CK+	78.65%

Table III comprehensively evaluates Model C, detailing precision, recall, F1-score, and support for each emotion class on the FER2013 dataset.

TABLE V. CLASSIFICATION METRICS FOR MODEL C ON FER2013 DATASET

Emotion	Precision	Recall	F1-Score	Support
Angry	0.62	0.62	0.62	491
Disgust	0.70	0.80	0.75	55
Fear	0.59	0.50	0.54	528
Happy	0.92	0.89	0.90	879
Sad	0.66	0.76	0.70	626
Surprise	0.59	0.58	0.59	594
Neutral	0.78	0.81	0.79	416

Model C performed much better, with an accuracy of 71.23% on FER2013, and outperformed models A and B. This improvement is due to the improved dropout and layer-wise strategy in Model C. On the CK+ dataset, Model A shows the best performance with an accuracy of 95.51%, followed by Model C with 78.65% and Model B with 83.15%. Since Model A performed so well on CK+, it indicates that a dropout-free model architecture does not hinder its performance. On the contrary, it may even help small, more controlled datasets like CK+, where overfitting is not as much of a problem.

B. Overfitting and Model Generalization

Overfitting occurs when a model learns the training data too well, capturing noise and specific patterns that do not generalize to unseen data. To evaluate the extent of overfitting and generalizability of these models, the accuracy and loss curves were used, as shown in Figures 3-5 for Models A, B, and C, respectively.

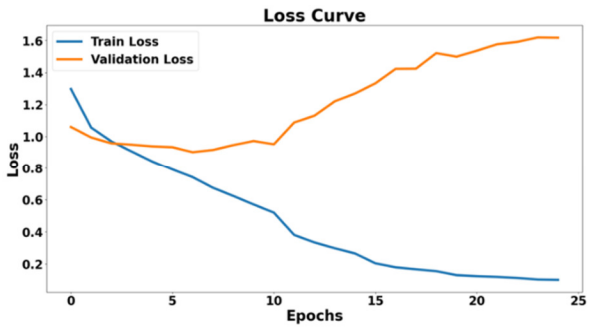

Fig. 3. Loss curves for Model A on FER2013.

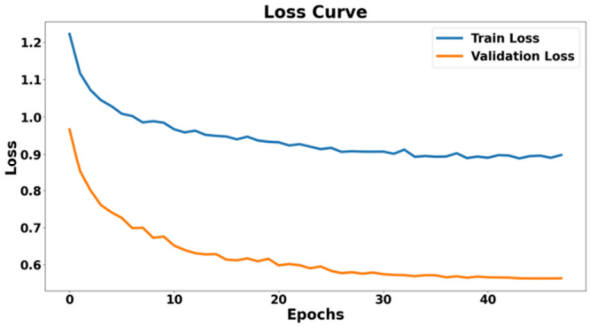

Fig. 4. Loss curves for Model B on FER2013.

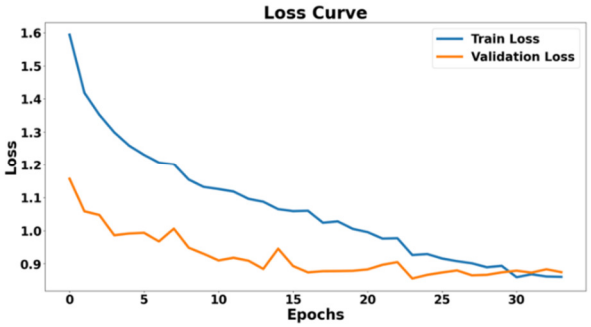

Fig. 5. Loss curves for Model C on FER2013.

Model C achieved a lower training loss and maintained a consistently lower validation loss throughout the training process, highlighting its effectiveness in minimizing overfitting. Model A, while achieving a low training loss, shows a slight increase in validation loss after initial epochs, indicative of overfitting. Model B maintained a balanced loss curve with gradual reductions in training and validation losses, reflecting its enhanced generalization performance through focal loss.

The architectural enhancements and training techniques employed in each model significantly influence their propensity to overfit and their ability to generalize:

- Model A: Utilizes a basic architecture without dropout, which allows it to achieve high training accuracy quickly. However, the absence of dropout leads to overfitting, as evidenced by the divergence between training and validation accuracy/loss curves.
- Model B: Incorporates focal loss, which helps address class imbalance by focusing more on hard-to-classify examples. This modification results in more stable validation performance and reduced overfitting compared to Model A, as reflected in its balanced accuracy and loss curves.
- Model C: Implements enhanced dropout and layer-wise unfreezing strategies. Enhanced dropout provides robust regularization, preventing the model from becoming overly dependent on specific training features. Layer-wise unfreezing allows for gradual fine-tuning of deeper layers, enabling the model to adapt to the dataset's nuances without disrupting pre-trained representations. These techniques collectively contribute to its superior generalization performance, as demonstrated by its high and stable validation accuracy and consistently low validation loss.

C. Model Performance

Confusion matrices visually represent the model's performance across different classes, highlighting areas where the model excels or struggles. Figure 6 presents the confusion matrices for Model A. The confusion matrix for Model B, shown in Figure 7, illustrates a more varied performance across different emotion classes.

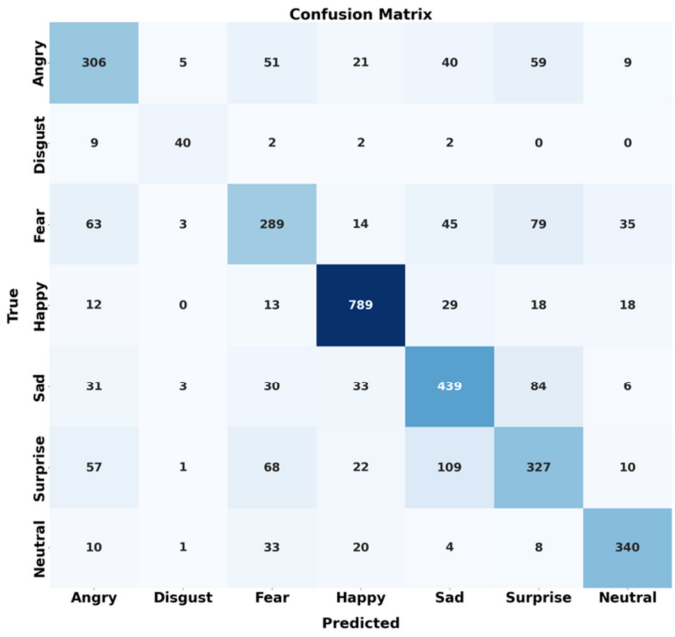

Fig. 6. Confusion matrix for Model A on FER2013.

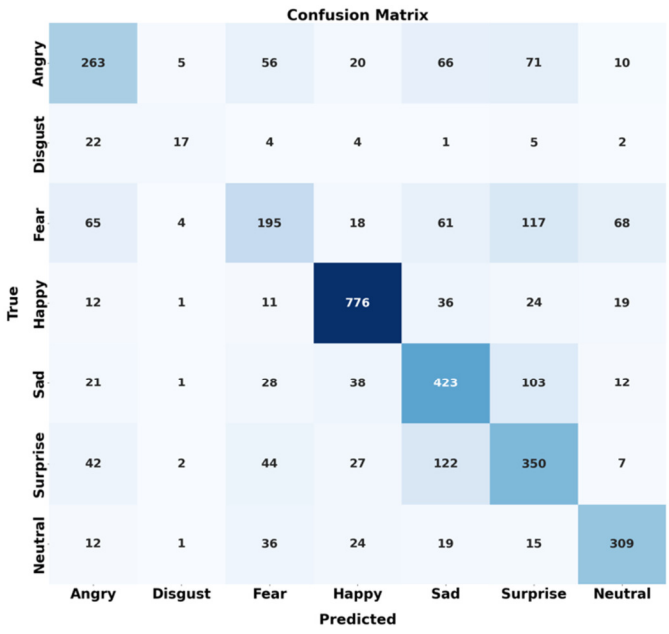

Fig. 7. Confusion matrix for Model B on FER2013.

The confusion matrix for Model C demonstrates a balanced performance across most emotion classes. The model achieves high precision and recall for the Happy and Neutral emotions, similar to Model A. Notably, Model C shows improved accuracy in recognizing Sad emotions compared to Models A and B. However, Fear and Surprise still present challenges, albeit to a lesser extent than Model B. Figure 8 displays the confusion matrix for Model C on the FER2013 dataset.

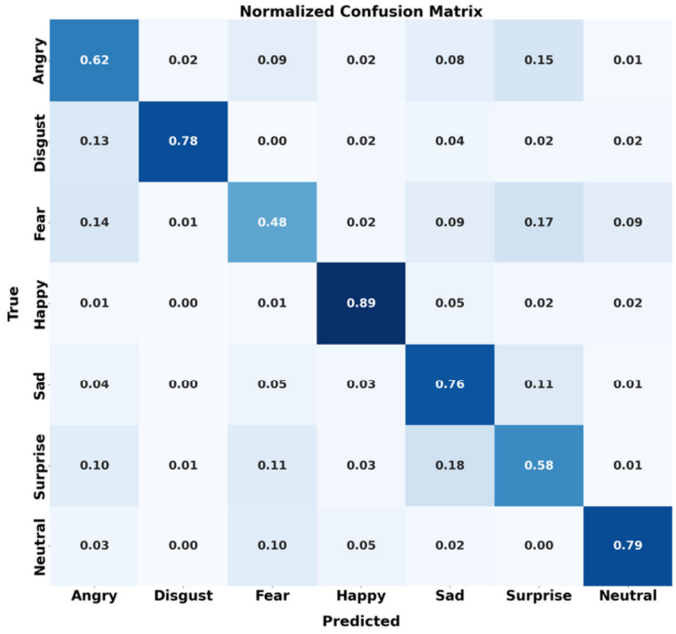

Fig. 8. Confusion matrix for Model C on FER2013.

D. Grad-CAM Visualizations

Figure 9 below presents Grad-CAM visualizations for Model C on selected FER2013 images, displaying both the original image and the corresponding Grad-CAM heatmap. The left image shows the original facial expression, while the right panel overlays the Grad-CAM heatmap on the original image.

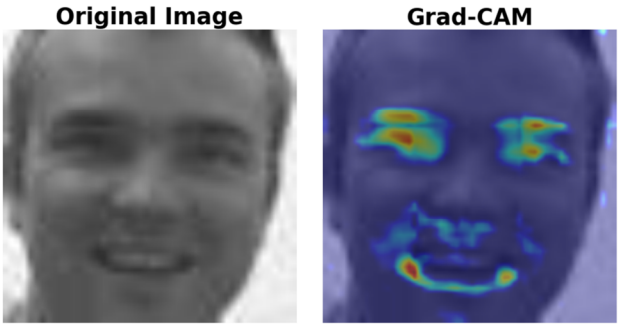

Fig. 9. Confusion matrix for Model C on FER2013.

The heatmap highlights the regions most influential in the model's prediction of each emotion:

- Happy: Intense focus on the mouth and eyes, capturing smiles and expressive eyes.
- Angry: Emphasis on the eyebrows and mouth area, indicating furrowed brows and tightened lips.
- Sad: Focus on the eyes and the downward turn of the mouth, reflecting subtle emotional cues.
- Neutral: Evenly distributed activation throughout the face, indicating a balanced assessment.

E. Mental Health Scoring

The mental health scoring system was developed by applying Model C to employee facial data collected over 30 days using the Genius HR system. The model predicted daily emotions from attendance images, and these predictions were averaged over 30 days to calculate MS for each employee. MS was derived using the interpolation formula, where higher scores indicate more significant mental health risks. The process included three key steps:

- Emotion prediction: Model C generated probability distributions for each emotion across all images, with daily emotion averages calculated over 30 days.
- Confidence scoring: The model's confidence levels for each prediction were tracked, and an average confidence score was calculated for each employee across the 30 days.
- Mental health score calculation: The MS was derived using the formula.

Table VI summarizes the mental health scores for the employees, calculated based on the average predicted emotions and the confidence scores. For example, Employee 39 had a mental health score of 53.00, reflecting elevated risks due to consistently negative emotions. In contrast, Employee 15 had a lower score of 50.93, indicating a more stable emotional state.

TABLE VI. MENTAL HEALTH SCORING SUMMARY

ID	Avg confidence scores	Number of images	MS score
31	0.7747	30	52.03
39	0.9230	30	53.00
16	0.8943	30	53.00
15	0.6484	30	50.93
17	0.7503	30	51.07

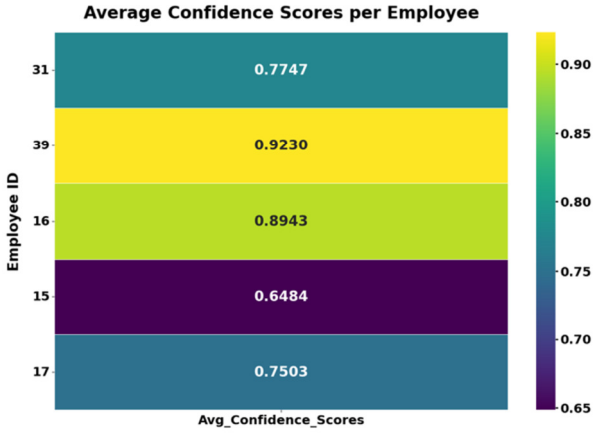

Fig. 10. Heatmap for average confidence scores.

Figure 10 shows the average confidence scores per Employee for 30 days, allowing the identification of those instances where the model was most confident in its emotional predictions. Confidence scores ranged from 0.6484 to 0.9230, with Employee 39 having the highest average of 0.9230, while Employee 15 had the lowest with 0.6484. These scores are related to the regularity of emotion forecasting that the model performed. Although the higher the confidence, as it was for Employee 1, the more regularly the model classified emotions, clear patterns of emotions were present. In turn, a small confidence value for Employee 15 should imply more uncertainty or ambiguity in how emotions were detected.

IV. CONCLUSION

This study presented a facial expression recognition model based on the Swin transformer-based FER model optimized for mental health detection. A layer-by-layer unfreezing strategy was used to maximize dropout and reduce overfitting. Model C surpassed the other proposed models, with accuracies of 71.23% on FER2013 and 78.65% on CK+. The weighted cross-entropy loss was used to balance the classes, which was enhanced by the proposed augmentation data approach to ensure the robustness of the result. This model was tested with Genius HR data for 30 days. Mental health ratings generated using Model C were between 50.93 and 53.00, reflecting that the model could help determine at-risk workers. The results demonstrate Swin transformers for efficient FER performance in diagnosing mental health disorders. Future research should improve the model's sensitivity to subtle emotions in diverse scenarios.

The primary novelty of this study lies in integrating enhanced dropout and layer-wise unfreezing strategies within the Swin transformer architecture for FER tasks, explicitly tailored for mental health detection. Although previous studies have employed Swin transformers for FER [19, 20, 24], none have combined these overfitting mitigation techniques in this context. Finally, this approach addresses overfitting, which is evident by comparing Model C's superior generalization performance to Models A and B. Furthermore, this study provides a practical use case for integrating the FER model with a mental health scoring system, which has not been performed or taken deep in most related works.

Compared to similar studies, such as [19], that utilized Swin transformers for FER without focusing on overfitting reduction, this model achieved superior performance through the proposed enhancements. Furthermore, in [24], fine-tuned Swin transformers were used for FER but did not incorporate a layer-wise unfreezing strategy or apply the model to mental health detection. This work bridges this gap by improving FER accuracy and demonstrating the model's utility in real-world mental health assessment scenarios. In conclusion, this study contributes to the field by presenting an innovative FER model that combines advanced overfitting mitigation techniques with the Swin transformer architecture and applies it to mental health detection, offering a novel framework for early workplace mental health intervention.

REFERENCES

[1] A. Malik *et al.*, "Mental health at work: WHO guidelines," *World Psychiatry*, vol. 22, no. 2, pp. 331–332, 2023, https://doi.org/10.1002/wps.21094.

[2] J. Aina, O. Akinniyi, Md. M. Rahman, V. Odero-Marah, and F. Khalifa, "A Hybrid Learning-Architecture for Mental Disorder Detection Using Emotion Recognition," *IEEE Access*, vol. 12, pp. 91410–91425, 2024, https://doi.org/10.1109/ACCESS.2024.3421376.

[3] S. Minaee, M. Minaei, and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *Sensors*, vol. 21, no. 9, Apr. 2021, Art. no. 3046, https://doi.org/10.3390/s21093046.

[4] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215, Jul. 2022, https://doi.org/10.1109/TAFFC.2020.2981446.

[5] A. A. A. Al-zanam, O. J. A. E. H. Alhomery, and C. P. Tan, "Mental Health State Classification Using Facial Emotion Recognition and Detection," *International Journal on Advanced Science Engineering Information Technology*, vol. 13, no. 6, pp. 2274–2281, 2023.

[6] S. M. Hassan, A. Alghamdi, A. Hafeez, M. Hamdi, I. Hussain, and M. Alrizq, "An Effective Combination of Textures and Wavelet Features for Facial Expression Recognition," *Engineering, Technology & Applied Science Research*, vol. 11, no. 3, pp. 7172–7176, Jun. 2021, https://doi.org/10.48084/etasr.4080.

[7] M. Mujiyanto, A. Setyanto, E. Utami, and K. Kusrini, "Facial Expression Recognition with Deep Learning and Attention Mechanisms: A Systematic Review," in *2024 7th International Conference on Informatics and Computational Sciences (ICICoS)*, Semarang, Indonesia, Jul. 2024, pp. 12–17, https://doi.org/10.1109/ICICoS62600.2024.10636857.

[8] P. Jiang, G. Liu, Q. Wang, and J. Wu, "Accurate and Reliable Facial Expression Recognition Using Advanced Softmax Loss With Fixed Weights," *IEEE Signal Processing Letters*, vol. 27, pp. 725–729, 2020, https://doi.org/10.1109/LSP.2020.2989670.

[9] R. Vedantham, "Adaptive increasing-margin adversarial neural iterative system based on facial expression recognition feature models," *Multimedia Tools and Applications*, vol. 81, no. 3, pp. 3793–3830, Jan. 2022, https://doi.org/10.1007/s11042-021-11320-1.

[10] Y.-J. Xiong, Q. Wang, Y. Du, and Y. Lu, "Adaptive graph-based feature normalization for facial expression recognition," *Engineering Applications of Artificial Intelligence*, vol. 129, Mar. 2024, Art. no. 107623, https://doi.org/10.1016/j.engappai.2023.107623.

[11] Z. Sun, C. Fu, M. Luo, and R. He, "Self-Augmented Heterogeneous Face Recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, Shenzhen, China, Aug. 2021, pp. 1–8, https://doi.org/10.1109/IJCB52358.2021.9484335.

[12] L. Wang, X. Kang, F. Ding, S. Nakagawa, and F. Ren, "A joint local spatial and global temporal CNN-Transformer for dynamic facial expression recognition," *Applied Soft Computing*, vol. 161, Aug. 2024, Art. no. 111680, https://doi.org/10.1016/j.asoc.2024.111680.

[13] Y. Liu, "Deep Learning-Driven Real-Time Facial Expression Tracking and Analysis in Virtual Reality," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Jan. 2024, Art. no. 20242283, https://doi.org/10.2478/amns-2024-2283.

[14] A. Barman and P. Dutta, "Facial expression recognition using Reversible Neural Network," *Applied Soft Computing*, vol. 162, Sep. 2024, Art. no. 111815, https://doi.org/10.1016/j.asoc.2024.111815.

[15] H. V. Manalu and A. P. Rifai, "Detection of human emotions through facial expressions using hybrid convolutional neural network-recurrent neural network algorithm," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, Art. no. 200339, https://doi.org/10.1016/j.iswa.2024.200339.

[16] J. Zhang, W. Wang, X. Li, and Y. Han, "Recognizing facial expressions based on pyramid multi-head grid and spatial attention network," *Computer Vision and Image Understanding*, vol. 244, Jul. 2024, Art. no. 104010, https://doi.org/10.1016/j.cviu.2024.104010.

[17] E. S. Agung, A. P. Rifai, and T. Wijayanto, "Image-based facial emotion recognition using convolutional neural network on emognition dataset,"

Scientific Reports, vol. 14, no. 1, Jun. 2024, Art. no. 14429, https://doi.org/10.1038/s41598-024-65276-x.

[18] X. Chen, X. Zheng, K. Sun, W. Liu, and Y. Zhang, "Self-supervised vision transformer-based few-shot learning for facial expression recognition," *Information Sciences*, vol. 634, pp. 206–226, Jul. 2023, https://doi.org/10.1016/j.ins.2023.03.105.

[19] M. Bie, H. Xu, Y. Gao, K. Song, and X. Che, "Swin-FER: Swin Transformer for Facial Expression Recognition," *Applied Sciences*, vol. 14, no. 14, Jul. 2024, Art. no. 6125, https://doi.org/10.3390/app14146125.

[20] A. Vats and A. Chadha, "Facial Expression Recognition using Squeeze and Excitation-powered Swin Transformers." arXiv, Apr. 29, 2023, https://doi.org/10.48550/arXiv.2301.10906.

[21] A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang, "DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–15, 2022, https://doi.org/10.1109/TIM.2022.3178991.

[22] L. Qin *et al.*, "SwinFace: A Multi-Task Transformer for Face Recognition, Expression Recognition, Age Estimation and Attribute Estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2223–2234, Apr. 2024, https://doi.org/10.1109/TCSVT.2023.3304724.

[23] S. Han, H. Chang, Z. Shi, and S. Hu, "Facial Expression Recognition Algorithm Based on Swin Transformer," in *2023 9th International Conference on Systems and Informatics (ICSAI)*, Changsha, China, Dec. 2023, pp. 1–6, https://doi.org/10.1109/ICSAI61474.2023.10423327.

[24] H. Feng, W. Huang, D. Zhang, and B. Zhang, "Fine-Tuning Swin Transformer and Multiple Weights Optimality-Seeking for Facial Expression Recognition," *IEEE Access*, vol. 11, pp. 9995–10003, 2023, https://doi.org/10.1109/ACCESS.2023.3237817.

[25] Y. Wu, A. Xiong, J. Lai, J. Liang, and J. Chen, "DFF: Deformable Attention Transformer-Based with Facial Feature Fusion Network for Facial Express Recognition," in *2023 IEEE International Conference on Unmanned Systems (ICUS)*, Hefei, China, Oct. 2023, pp. 984–989, https://doi.org/10.1109/ICUS58632.2023.10318324.

[26] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin, "Cross-Domain Facial Expression Recognition: A Unified Evaluation Benchmark and Adversarial Graph Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9887–9903, Dec. 2022, https://doi.org/10.1109/TPAMI.2021.3131222.

[27] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

[28] N. Li, Y. Huang, Z. Wang, Z. Fan, X. Li, and Z. Xiao, "Enhanced Hybrid Vision Transformer with Multi-Scale Feature Integration and Patch Dropping for Facial Expression Recognition," *Sensors*, vol. 24, no. 13, Jan. 2024, Art. no. 4153, https://doi.org/10.3390/s24134153.

[29] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and Improving Relative Position Encoding for Vision Transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 10013–10021, https://doi.org/10.1109/ICCV48922.2021.00988.

[30] F. Scala, A. Ceschini, M. Panella, and D. Gerace, "A General Approach to Dropout in Quantum Neural Networks," *Advanced Quantum Technologies*, Art. no. 2300220, https://doi.org/10.1002/qute.202300220.

[31] I. J. Goodfellow *et al.*, "Challenges in Representation Learning: A Report on Three Machine Learning Contests," in *Neural Information Processing*, pp. 117–124, https://doi.org/10.1007/978-3-642-42051-1_16.

[32] J. Yang, Z. Lv, K. Kuang, S. Yang, L. Xiao, and Q. Tang, "RASN: Using Attention and Sharing Affinity Features to Address Sample Imbalance in Facial Expression Recognition," *IEEE Access*, vol. 10, pp. 103264–103274, 2022, https://doi.org/10.1109/ACCESS.2022.3210109.

[33] F. Xue, Q. Wang, Z. Tan, Z. Ma, and G. Guo, "Vision Transformer With Attentive Pooling for Robust Facial Expression Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3244–3256, Jul. 2023, https://doi.org/10.1109/TAFFC.2022.3226473.

[34] O. S. Ekundayo and S. Viriri, "Facial Expression Recognition: A Review of Trends and Techniques," *IEEE Access*, vol. 9, pp. 136944–136973, 2021, https://doi.org/10.1109/ACCESS.2021.3113464.

[35] G. Simcock *et al.*, "Associations between Facial Emotion Recognition and Mental Health in Early Adolescence," *International Journal of Environmental Research and Public Health*, vol. 17, no. 1, Jan. 2020, Art. no. 330, https://doi.org/10.3390/ijerph17010330.