Leveraging Large Language Models for the Identification of Human Emotional States

Clement Leung
School of Science and Engineering and
Guangdong Provincial Key Laboratory of
Future Networks of Intelligence
Chinese University of Hong Kong
Shenzhen, China
clementleung@cuhk.edu.cn

Zhifei Xu
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
zhifeixu1@link.cuhk.edu.cn

***Abstract*—This study explores emotion recognition, classification, and prediction, emphasizing its importance in safety-critical tasks where emotional competence can save lives. We classify emotions into positive (competent) and negative (incompetent) states and develop a stochastic model featuring an emotion stability factor, to measure how quickly individuals return to their baseline emotional state—lower indicates greater emotional stability. Our model provides a foundation for personalized emotional health strategies and tailored psychological treatments. Additionally, we evaluate ChatGPT-4's zero-shot capabilities in image-based sentiment reasoning compared to ResNet-50 and Vision Transformer models. Despite competitive performance, challenges such as unstable predictions underscore the complexity of mental health analysis in image conversations. We propose improvements through enhanced prompt engineering, model fine-tuning, and an ensemble approach combining each model's strengths to create a more accurate emotion classification system with significant implications for mental health applications.**

Keywords-image emotion recognition; large language model; zero-shot; emotion stability factor; ChatGPT4.

I. Introduction

The complex nature of human interactions makes it essential to accurately perceive and express emotions. Emotions shape individual experiences, influencing interactions, decision-making, and overall well-being. They help build connections and act as indicators of personal intentions and health.

For over a decade, research has focused on integrating emotional responsiveness into human-computer dialogue systems [1][2][3][4][5][6][7]. This field emphasizes the importance of understanding and predicting emotions for enhancing human-computer interactions and broader applications. As technology advances, there is a need to develop systems that can recognize and react to a wide range of emotions, ensuring natural interactions.

In modern society, individuals face stress from criticism, injustice, or relationship issues, often leading to emotional instability. Research shows that stress can result in harmful behaviors [8][9]. Such emotional distress can have severe consequences, from academic pressures leading to suicidal thoughts to road rage incidents. In high-stakes jobs like aviation or surgery, emotional regulation is critical for safety. Recent incidents, such as a pilot attempting to shut down an engine mid-flight due to depression, highlight the importance of emotional assessment. Key reasons why emotion recognition is important include:

- Understanding Emotions: Emotions guide how we perceive situations and decide what we want. Recognizing emotions helps individuals make choices that lead to positive interactions.
- Impact of Emotions: Emotions affect moods and behaviors, influencing relationships and well-being. Recognizing when we feel unhappy allows us to find solutions and regain control.
- Negative Emotions and Thoughts: Failing to recognize negative emotions can lead to harmful thoughts. By changing negative thinking patterns, individuals can reshape their perspectives and improve well-being.

This dissertation explores emotion recognition, classification, and prediction in high-stakes environments. We categorize emotional states into positive (competent) and negative (incompetent) groups. Our stochastic model analyzes how emotions evolve, using the emotional stability factor (λ) to quantify how quickly emotions return to a baseline. A lower λ indicates greater emotional stability, while a higher λ suggests frequent emotional fluctuations.

Understanding these dynamics can enhance therapeutic interventions, making them more effective and personalized. By anticipating emotional declines, the model enables the design of timely interventions. Knowing one's emotional stability factor is crucial for managing mental health, building resilience, and improving quality of life.

Section II reviews advancements in neural networks, including their use in online social networks and deep learning technologies. Models like ChatGPT [10] and Instruct-GPT [11] demonstrate the potential for improving emotion recognition.

This study investigates ChatGPT4's zero-shot capability to recognize emotions from images. While promising, ChatGPT4 has limitations, including prediction instability and reasoning inaccuracies. Enhancing its effectiveness through better prompt engineering and context selection is crucial for its application in mental health assessment. Our main contributions include:

- Applying mathematical models to predict emotional changes, offering a foundation for personalized mental health strategies.
- Using different emotional stability factors to understand how environments, individuals, or systems return to baseline states, aiding the design of targeted interventions.

- Exploring ChatGPT4's potential for emotion recognition in zero-shot scenarios.
- Developing conversational techniques in ChatGPT4 to enhance emotion recognition and discussing progress and limitations in its multimodal tasks [12][13][14].

The rest of this paper is organized as follows. Section II reviews the development of emotion recognition research and related studies. Section III explores the theoretical foundations of emotion recognition and prediction through rigorous mathematical logic and formulas. It lists the derivations of the algorithms and methods adopted in emotion recognition systems. By providing a solid mathematical framework, this section ensures that the subsequent empirical analysis is based on a robust theoretical model, which helps to gain a deeper understanding of the mechanisms behind emotion recognition technology. Section IV provides an experimental study focused on emotion recognition. It first details the selection process of appropriate datasets that are representative and comprehensive, thereby ensuring the validity and reliability of the experiments. This section further defines the specific tasks formulated as part of the research and outlines the goals and expected results of these experiments. The experimental results are analyzed, providing insights into the effectiveness of current emotion recognition models and highlighting potential areas for improvement. Section V and VI combine the experimental results with the capabilities of ChatGPT4 in the field of emotion recognition and prediction. It evaluates the progress brought by ChatGPT4 and discusses the performance of the model in the context of the experimental results. The paper ends with a reflection on the significance of these findings for future research and practical applications in this field, and provides suggestions for further research and enhancement of existing models.

II. Related work

What is emotion recognition? It refers to the technological process of identifying and interpreting human emotions. People naturally vary in their ability to accurately recognize others' feelings, which has led to the development of a specialized field that leverages technology to assist in this task. A key concept in this domain is affective forecasting, also known as hedonic forecasting. Affective forecasting involves predicting one's future emotional states, which plays a crucial role in shaping individual preferences, decisions, and behaviors. This interdisciplinary field, studied by both psychologists and economists, has a wide range of applications across various sectors.

The formal study of emotion theory began with Charles Darwin in 1872, establishing a foundation for future research into emotional expression. Building on Darwin's work, Paul Ekman introduced one of the most influential classification models in emotion recognition. He identified six basic emotions—joy, sadness, fear, anger, surprise, and disgust—which he proposed as universally recognizable across various cultures [15].

Robert Plutchik further expanded upon Ekman's model by developing the "wheel of emotions," which includes eight primary emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation [16]. Plutchik's wheel illustrates how these basic emotions can blend to form more complex emotional experiences, highlighting the dynamic nature of human emotions.

The theoretical frameworks for classifying emotions generally fall into two categories: categorical models and continuum models. Categorical models, or discrete emotion models, are based on the premise that there are a finite number of primary or basic emotions that are universally experienced, regardless of cultural or individual differences. In contrast, the continuum model views emotions as existing along a spectrum, capturing the nuances of emotional experience through dimensions such as valence (positive to negative), arousal (excited to calm), and dominance (sense of control or influence in a situation) [17] [18].

In recent years, emotion recognition and prediction technologies have developed into two primary types: single-modal and multimodal systems. Single-modal systems rely on a single data type, such as text, speech, or facial expressions, to detect and predict emotions [19] [20]. However, these systems often face limitations due to the constrained information provided by a single data source. For instance, relying solely on facial expressions may not fully capture an individual's emotional state, particularly in complex scenarios. The technology behind emotion recognition and prediction has gained significant attention due to its potential to enhance safety, support mental health, and improve user experiences. It has been identified as a vital factor in promoting human safety, making it a focus of extensive research [8][9]. The significance of emotion recognition and prediction is further highlighted by the growth of the Emotion Detection and Recognition (EDR) market. In 2024, the EDR market was valued at $57.25 billion and is projected to reach $139.44 billion by 2029, reflecting a remarkable compound annual growth rate (CAGR) of 19.49% from 2024 to 2029. This rapid expansion signifies the increasing demand and versatility of emotion recognition technology across various sectors. The growth is driven by the technology's ability to provide valuable insights into human emotions, proving invaluable in areas such as marketing, customer service, therapy, and security. By accurately detecting and responding to human emotions, this technology not only enhances interactions but also contributes to overall safety and well-being [21].

III. Methods

In real-world settings, emotions are dynamic and constantly shift from one state to another, which is crucial to consider in operational environments like workplace scheduling or hospital management. Accurately predicting individuals' emotional states is vital, especially when assigning tasks that demand high concentration and emotional stability. This need is even more pronounced in safety-critical roles. Workers experiencing emotional distress, whether due to unfair treatment or fatigue, may not only underperform but also pose safety risks to themselves and others. Thus, ensuring that employees, particularly in high-risk industries, maintain emotional stability is essential for workplace safety and efficiency.

Assessing emotional states is key to preventing potential accidents or errors that can arise from impaired judgment caused by negative emotions. For example, a worker struggling with unmanaged anger or severe sadness may lack the focus or decision-making capacity needed to safely operate machinery or make critical, split-second decisions. Therefore, understanding and managing employees' emotional well-being goes beyond enhancing individual performance; it is also about protecting the overall safety and smooth functioning of the workplace.

We apply Ekman's model, which identifies six primary emotions: happiness, sadness, fear, anger, surprise, and disgust. In this model, we categorize happiness and surprise as positive $+1$ emotions, suggesting a state of emotional well-being and openness. On the other hand, sadness, fear, anger, and disgust are categorized as negative -1 emotions, which might indicate potential emotional instability. Notably, in environments where safety is paramount, such as high-risk jobs, the categorization of surprise may shift. Typically considered a positive emotion due to its association with unexpected joy, surprise can be reclassified as negative in these critical contexts to ensure a conservative approach to emotional management, thereby reducing the risk of abrupt emotional reactions that could impair judgment or performance. Then, we propose that a crucial element in this model is the emotion stability factor, denoted by λ. Individuals with smaller emotion stability factor values exhibit more consistent emotional states. They are able to sustain either a positive or a neutral mood over extended periods, indicating resistance to sudden changes in emotions caused by external circumstances or internal reflections. Conversely, individuals with higher emotion stability factor values are prone to frequent emotional fluctuations. This heightened emotional reactivity can be attributed to various causes, including external influences like social interactions or internal factors such as hormonal variations. We will explore this concept in further detail in the sections that follow.

We represent the emotional state at time t by $S(t)$, where t denotes time and $S(t)$ describes the individual's emotional changes over time. The emotional state $S(t)$ can take on the following values:

$$S(t) = +1 \tag{1}$$

Equation (1) corresponds to the person's emotion being positive,

$$S(t) = -1 \tag{2}$$

Equation (2) means that the person's emotion is negative, corresponding to positive $+1$ and negative -1 emotions. Since various external happenings continually bombard humans, mood changes are often caused by events outside their control, possibly due to various factors. Such factors may be related to changing conditions of financial situation, relationships, health, work, stock market, and family, and the combination of these may cause a transition from a positive emotion state to a negative emotion state and vice versa.

First, let $S(0) = 1$. Then, we represent the transition time points by a Poisson Process. Now, $S(t) = 1$ if the number of transitions in the time interval $(0, t)$ is even, and $S(t) = -1$ if this number is odd. Therefore,

$$P[S(t) = 1|S(t) = 0] = p_0 + p_2 + p_4 + ... + ..., \tag{3}$$

where p_k is the number of Poisson points in $(0, t)$ with parameter λ:

$$p_k = e^{-\lambda t}\frac{(\lambda t)^k}{k!} \tag{4}$$

That is,

$$\begin{aligned} P[S(t) = 1|S(0) = 0] &= e^{-\lambda t}[1 + \frac{(\lambda t)^2}{2!} + \frac{(\lambda t)^4}{4!}... + ...] \\ &= e^{-\lambda t}\cosh \lambda t \end{aligned} \tag{5}$$

with

$$\cosh(t) = \frac{e^{-t} + e^{t}}{2} \tag{6}$$

$\cosh(t)$ can be expressed by its series expansion:

$$\cosh(t) = \sum_{k=0}^{\infty}\frac{t^{2k}}{(2k)!} \tag{7}$$

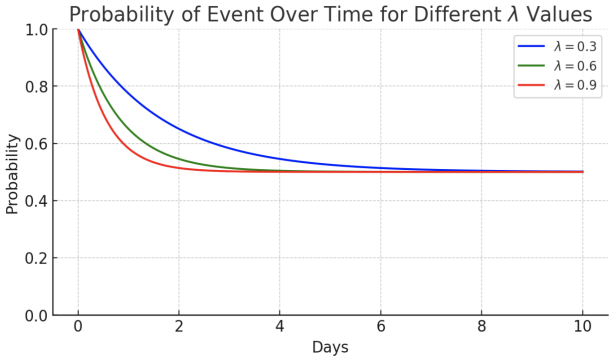

Figure 1. Different Emotion Stability Factors of the Equation (5)

On the other hand, Figure 1 provides a valuable perspective on how individuals might experience shifts in emotional states under different conditions. For example, individuals with a lower emotional stability factor, such as $\lambda = 0.3$, could be likened to those with a strong emotional foundation or support system, allowing them to maintain a higher probability of positive emotional states for longer periods. This may reflect traits like resilience, optimism, or effective coping mechanisms that slow the rate of emotional decay after positive experiences.

In contrast, higher emotional stability factor values, such as $\lambda = 0.9$, might represent individuals who are more susceptible to rapid emotional shifts due to external stimuli or internal factors, such as stress or a lack of coping resources. These individuals experience a quicker return to baseline emotional states, making them more vulnerable to negative emotional swings following positive experiences.

Examining the day-to-day probabilities reveals that achieving and maintaining a probability of over 50% for positive emotions

within the first day is feasible across all values of the emotional stability factor, with probabilities ranging from 60% to 80%. However, the duration over which this level is maintained varies significantly. While an emotional stability factor of $\lambda = 0.3$ can sustain this level for up to four days, an emotional stability factor of $\lambda = 0.9$ sees a rapid decrease, reaching 50% by the second day. This rapid decline may indicate the need for more frequent or stronger positive reinforcements or interventions to maintain positivity in such individuals.

Additionally, the stabilization of probabilities around the 50% mark from the second day for higher emotional stability factors and from the fourth day for lower factors suggests a natural equilibrium state in emotional dynamics. This equilibrium represents a critical point where the emotional state is equally likely to shift towards positive or negative, emphasizing the importance of timely psychological support or self-care practices during these periods to tilt the balance towards a more positive state.

In practical terms, this analysis can help tailor therapeutic approaches or wellness programs to fit individual emotional profiles. Understanding these dynamics could lead to more personalized and effective interventions, enhancing emotional well-being and stability by strategically addressing the decay rates of positive emotional states. Such insights are invaluable in clinical psychology, counseling, and personal development, where maintaining positive emotional states is essential for mental health and quality of life.

Now, if $S(t) = -1$ when the number of points in the time interval $(0, t)$ is odd, we have:

$$P[S(t) = -1|S(0) = 0]] = e^{-\lambda t}[1 + \frac{(\lambda t)^3}{3!} + \frac{(\lambda t)^5}{5!} ... + ...] = e^{-\lambda t} \sinh \lambda t \quad (8)$$

and

$$\sinh(t) = \frac{e^t - e^{-t}}{2} \quad (9)$$

This series is equivalent to the series expansion of $\sinh(t)$:

$$\sinh(t) = \sum_{k=0}^{\infty} \frac{t^{2k+1}}{(2k+1)!} \quad (10)$$

Equation (5) represents the probability that the emotion remains positive at time t, given that it was positive at time 0. Similarly, Equation (8) provides the probability that the emotion is positive at time t, given that it was negative at time 0. In both expressions, the emotional stability factor determines the rate of emotional change or decay over time. A larger value of this factor indicates more rapid emotional shifts, while a smaller value suggests slower changes.

Figure 2 illustrates the function $e^{-\lambda t} \sinh(\lambda t)$ with emotion stability factor values of 0.3, 0.6, and 0.9, offering insights into how probability changes over time, particularly in the context of emotional states or other processes that evolve or decay similarly. The graph shows that different values of the emotional stability factor (λ) can significantly affect the duration and intensity of these probability states over days.

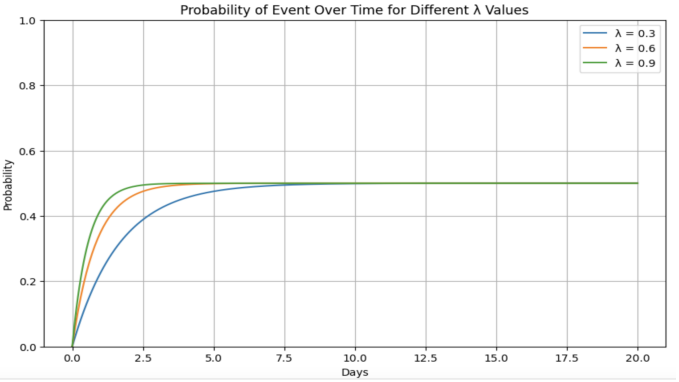

Figure 2. Different Emotion Stability Factors of the Equation (8)

For a lower decay constant, such as $\lambda = 0.3$, the probability starts strong and decays slowly, indicating a lingering effect. In the context of maintaining a negative emotional state, this suggests a higher likelihood of staying in that state for an extended period. Within the first day, the probability remains above 60%, indicating a persistent state that gradually stabilizes near 50% by the third day. This slower decay could reflect situations where the factors causing the emotional state are not quickly resolved.

When the decay constant increases to $\lambda = 0.6$, the probability declines more steeply, signifying a quicker dissipation of the state. While the probability is initially high, it drops below 60% by the end of the first day and approaches 50% by the second. This faster decay could represent scenarios where interventions are more effective or where individuals have better-coping mechanisms.

With a higher decay constant ($\lambda = 0.9$), the drop in probability is even more rapid. The steep descent shows that the state dissipates quickly, failing to sustain above 50% for more than two days and nearing this threshold by the end of the first day. This could represent situations where external support is highly effective, or the cause of the emotional state has a short-lived impact.

By analyzing these curves, we can infer the effectiveness of different strategies or inherent factors in managing specific emotional or physical states. The varying stability factor values represent different rates at which environments, individuals, or systems return to baseline or transition between states. Understanding these temporal dynamics is crucial in fields like psychology, where predicting the duration of a negative emotional state can inform tailored, time-sensitive interventions aligned with observed decay rates.

IV. Results

Emotion Recognition in Conversations (ERC) is extensively used in various contexts. This includes analyzing comments on social media platforms and monitoring personnel in high-stress environments. In addition, ERC technology is used in chatbots to accurately assess users' emotional states so they can tailor their responses accordingly. ChatGPT4 is one such conversational bot, as highlighted earlier. Within the context

of these interactions, we investigate and evaluate how well it recognizes and understands emotions and sentiments.

1) Dataset and Evaluation Graph: We using three different datasets from Kaggle, Facial Expressions Training Data [22], Emotion Detection [23], and Natural Human Face Images for Emotion Recognition [24].

Emotion Detection This dataset consists of 35,685 examples of 48x48 pixel grayscale images, which contain two folders, one is trained, and the other one is tested. The folders contain different categories of emotional images. In addition, the images have been labeled by the authors for different types of emotions, including anger, disgust, fear, happiness, neutral, sad, and surprise.

Facial Expressions Training Data AffectNet [25] is a large database of faces marked with "impact" (the psychological term for facial expressions). In order to accommodate common memory limitations in this dataset, the authors reduce the resolution to 96x96 for the neural network processing, which indicates that all images are 96x96 pixels. Meanwhile, using Singular Value Decomposition, each image's Principal Component Analysis is calculated. The threshold for the Percentage of the First Component (index 0) in the principal components (in short the PFC%) was set to lower than 90%. This means that most if not all of the monochromatic images were filtered out. Finally, the dataset is based on Affectnet-HQ, using a state-of-the-art Facial Expression Recognition (FER) model that refines the AffectNet original label to re-label its dataset, which contains eight emotional categories - anger, contempt, disgust, fear, happiness, neutral, sadness, and surprise.

Natural Human Face Images for Emotion Recognition Since facial expression recognition is usually performed using standard datasets, such as the Facial Expression Recognition dataset (FER), Extended Cohn-Kanade dataset (CK+) and Karolinska Directed Emotional Faces dataset (KDEF) for machine learning, however, this dataset was collected from the internet and manually annotated to provide additional data on real faces, with over 5,500 + images with 8 emotions categories: anger, contempt, disgust, fear, happiness, neutrality, sadness and surprise. All images contain grayscale human faces (or sketches). Each image is 224 x 224 pixel grayscale in Portable Network Graphics (PNG) format. Images are sourced from the internet where they are freely available for download e.g., Google, Unsplash, Flickr, etc.

A. Task Definition

Based on the dataset description outlined in the previous section, we selected six emotions that are consistently annotated in each dataset for our experiments. These emotions include anger, disgust, happiness, neutrality, sadness, and surprise. Table I illustrates several examples of comparing annotations with ChatGPT4's predictions, with differences highlighted in red. For our experimental setup, we randomly selected 50 images representing six emotion types from each dataset and then submitted these images to ChatGPT4 for evaluation. Given that ChatGPT4 was launched in 2023, all of our experiments used this version of the model. At the same time, we chose to use the classic ResNet50 and ViT models as a comparison with ChatGPT.

ResNet50 is a classic deep convolutional neural network proposed by He Kaiming et al. [26], which effectively solves the gradient vanishing and degradation problems in deep networks by introducing residual connections. In the task of emotion recognition, subtle changes in facial expressions are crucial for accurate classification. With its powerful feature extraction capabilities, ResNet50 can effectively capture detailed features and complex patterns in face images. In our emotion recognition project, we used a pre-trained ResNet50 model as a feature extractor and then added a custom fully connected layer to adapt to the classification task of six emotion categories (anger, disgust, happiness, sadness, neutral, surprise). Through transfer learning, we made full use of the rich features learned by ResNet50 on large datasets such as ImageNet, which not only improved the accuracy of the model but also accelerated the training speed.

Vision Transformer (ViT) is a model proposed by the Google research team in 2020 that applies the Transformer architecture to image classification. Unlike traditional convolutional neural networks, ViT divides images into fixed-size image patches, which are then expanded into sequences and input into the Transformer network. ViT uses the self-attention mechanism to capture global dependencies in the image, which is very beneficial for understanding and classifying complex facial expressions. In our emotion recognition project, we took the pre-trained ViT model and fine-tuned it to adapt to the classification task of six emotion categories. The global feature capture capability of the ViT model gives it an advantage in identifying complex emotional expressions involving coordinated changes in multiple parts.

When evaluating the capabilities of ChatGPT4, we adopted a supervised learning approach and tested the performance of the model in a zero-shot prompting scenario designed specifically for this task. Each prediction of ChatGPT4 was carefully compared with our predefined cognitive assessment of the emotions depicted in the image. A correct prediction by ChatGPT4 that was consistent with our assessment was scored as 1, while a mismatch was scored as 0. In addition, each emotion was classified as positive, negative, or neutral according to ChatGPT4's description.

In addition, we generated a Receiver Operating Characteristic (ROC) curve based on the recorded results to quantify the accuracy of the model. When generating the ROC curve, emotions classified as positive (happy, neutral, or surprised) were marked with the factual result 1. In contrast, negative emotions (angry, disgusted, or sad) were marked with the factual result 0. The prediction accuracy of ChatGPT4 was then evaluated: if a positive emotion was correctly identified, it was recorded as 1; if not, it was recorded as 0. Similarly, for negative emotions, correct identification is recorded as 0, and incorrect identification is recorded as 1. In addition, the researchers also evaluated the confidence of each prediction using an evaluation

TABLE I
EXAMPLE OF CHATGPT4'S PREDICTION ON ERC TASK WITH IMAGES.

Image Content	Question	Annotation	Prediction
	What is the emotion of this person?	anger	surprise/shock/fear
	What is the emotion of this person?	happiness	happiness
	What is the emotion of this person?	happiness	happiness/joy
	What is the emotion of this person?	anger	frustration/concern/disapproval
	What is the emotion of this person?	sadness	sadness/crying
	What is the emotion of this person?	surprise	surprise

TABLE II
RESULT OF TRAINING RESNET50 MODEL PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

loss	accuracy	val_loss	val_accuracy
2.2056	0.163	327522.7813	0.1597
1.9918	0.2430	2906.5945	0.1528
1.7945	0.2205	770.0659	0.1528
1.7411	0.2466	518.6124	0.1667
1.6758	0.2812	36.9275	0.1597
1.6480	0.2951	2.2501	0.2569
1.6283	0.2795	2.8927	0.1875
1.6717	0.2830	7.9509	0.1944
1.5971	0.3056	11.4519	0.1667
1.5564	0.3698	4.6194	0.1875
1.5738	0.3072	8.1097	0.2013

index of 1 to 3 points, where 1 indicates low confidence, 2 indicates medium confidence, and 3 indicates high confidence. This structured evaluation helps to quantitatively evaluate the effectiveness of ChatGPT4 in identifying and distinguishing various emotional states based on facial expressions.

B. Results

ResNet-50 is a well-known convolutional neural network (CNN) architecture designed to tackle image classification tasks efficiently.We can see the training result of ResNet-50 in the Table II. In this emotion recognition context, it achieves an accuracy of 30.72%, indicating a moderate ability to capture visual patterns relevant to different emotional expressions. This performance can be attributed to ResNet-50's strong feature extraction capability, which results from its deep convolutional layers and residual connections. These connections enable the network to learn complex features from images while mitigating the vanishing gradient problem, which commonly plagues deep networks during training. The training history of ResNet-50 shows a relatively faster convergence rate compared to ViT, suggesting it can achieve an acceptable level of accuracy within fewer epochs. This attribute makes it useful in situations where computational resources or training time is limited.

The Vision Transformer introduces a novel approach to image classification by adopting the self-attention mechanism, which has proven highly successful in natural language processing (NLP) tasks. Unlike CNNs, ViT processes images as a series of patches, learning relationships between different parts of the image. In this context, ViT achieves a lower accuracy of 19.96% which show that result of training ViT in the Table III, suggesting it encounters challenges in optimizing for emotion recognition. However, the training history reveals a more stable validation loss trajectory compared to ResNet-50, indicating consistent learning and potential for improvement.

Unlike ResNet-50, which excels in extracting local features, ViT captures global relationships across the entire image through its self-attention mechanism. This global context is

TABLE III
RESULT OF TRAINING VIT MODEL PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

loss	accuracy	val_loss	val_accuracy
4.2872	0.1701	4.5549	0.1458
4.0688	0.1719	4.3315	0.1458
3.8694	0.1649	4.1312	0.1528
3.6925	0.1667	3.9536	0.1597
3.5350	0.1719	3.7932	0.1736
3.3960	0.1684	3.6490	0.1876
3.275	0.1667	3.5185	0.1667
3.1628	0.1649	3.3998	0.1806
3.0655	0.1667	3.2949	0.1806
2.9799	0.1719	3.1971	0.1806
2.9024	0.1701	3.1120	0.1806
2.8348	0.1788	3.0369	0.1806
2.7766	0.1788	2.9686	0.1806
2.7244	0.1927	2.9079	0.2014
2.6777	0.1910	2.8552	0.1945
2.6396	0.1910	2.8048	0.2084
2.6023	0.1910	2.763	0.2153
2.5710	0.1927	2.726	0.2153
2.5418	0.196	2.6929	0.2153
2.5176	0.1997	2.6595	0.2153

essential for emotion recognition, as emotions often depend on the overall configuration of facial features rather than isolated parts. The relatively stable validation loss in the training history of ViT suggests it learns in a more consistent manner. This stability could lead to better generalization with sufficient data and extended training. While the current accuracy is lower, its learning behavior indicates room for enhancement with more epochs, larger datasets, or additional fine-tuning. ViT can handle various input sizes and is not as constrained by fixed kernel sizes as CNNs, providing more flexibility when processing complex visual information such as varying facial expressions.

TABLE IV
RESULT OF CHATGPT4'S PREDICTION ON SINGLE EMOTION RECOGNITION TASK WITH IMAGES.

Emotion	**Accuracy**
Anger	30%
Disgust	19.30%
Happiness	78%
Neutral	69.34%
Sadness	44.30%
Surprise	70%

Table IV presents ChatGPT-4's performance in recognizing various single emotions. For the positive emotion of surprise, the model achieves an accuracy of approximately 70%. Meanwhile, the accuracy for identifying happiness reaches about 78%, highlighting ChatGPT-4's strong ability to detect positive emotions. When these two emotions are combined, the model maintains a commendable accuracy. As discussed in this section, both happiness and surprise correspond to a lower emotional stability factor, indicating that individuals can sustain these positive emotional states for extended periods.

In terms of negative emotions, ChatGPT-4's accuracy ranges from lowest to highest for disgust, anger, and sadness. During testing, we found that while the model can predict negative emotions in a zero-shot setting, it struggles to accurately differentiate between disgust and anger. The lowest accuracy is observed in identifying disgust, which may be attributed to the inconsistency in individual expressions of this emotion. Overall, ChatGPT-4's recognition accuracy from highest to lowest across the six emotions is as follows: happiness, surprise, neutral, fear, anger, and disgust. Notably, although ChatGPT-4 can correctly identify most images of surprise, it has difficulty determining whether the surprise is positive or negative, often categorizing it as a neutral emotion. This explains why the results for surprise closely resemble those for neutral.

As previously mentioned, mitigating the adverse effects of negative emotions is crucial, particularly for individuals in high-risk industries or groups. Therefore, our analysis focuses on three primary negative emotions: anger, disgust, and sadness, which are associated with a higher emotional stability factor and thus require greater attention. Our analysis reveals the following False Positive Rates (FPR): sadness at 0.3267, anger at 0.4800, and disgust at 0.6467. These FPR values indicate that disgust is the most challenging emotion to identify accurately, making it the most difficult category among the six evaluated emotions. The overall accuracy in detecting negative emotions remains insufficient. To address this, enhancing the prompts provided to ChatGPT-4 is essential. Although the model can recognize the presence of negative emotions in a zero-shot scenario, it struggles to accurately distinguish between specific states such as disgust, contempt, or anger. Therefore, more refined prompt engineering could improve the model's ability to discern these nuanced emotional states.

V. Discussion and Evaluation

The training history indicates that ResNet-50 starts with high and unstable validation loss, hinting at a potential overfitting risk. This suggests sensitivity to noise in the data, which may require careful hyperparameter tuning, regularization, and augmentation techniques to improve generalization. While its 30.72% accuracy shows some effectiveness, it remains significantly lower than GPT-4's 50.99%. This lower performance indicates that ResNet-50 might struggle to capture the holistic context of emotions, focusing instead on local features without considering global facial patterns.

The initial high training loss and the current low accuracy of 19.96% suggest that ViT requires more epochs and possibly larger datasets to reach optimal performance. This slower convergence is often a trade-off for its global feature extraction capability. ViT typically benefits from vast amounts of training data to learn complex patterns effectively. When applied to smaller datasets, such as those typically available for emotion recognition, it may struggle to outperform more established CNNs like ResNet-50 unless supplemented with data augmentation or transfer learning strategies.

During training, we often encounter inconsistencies between the images in specific datasets and our real-life perceptions. Since individuals react differently to images, biases can inadvertently be introduced into the emotional recognition of specific photos. For these particular images, we rely on human judgment as the ultimate arbiter, comparing our assessments with the outputs from ChatGPT4 to identify discrepancies and possible biases.

Furthermore, a significant challenge arises from the misalignment between the predictions of ChatGPT4 and the guidelines provided in the dataset. This divergence highlights a fundamental issue where ChatGPT4 often deviates from the dataset norms. For instance, if the dataset annotates an image depicting anger based on the subject's facial expression, ChatGPT4 might interpret the same expression as sadness or confusion. This difference does not necessarily imply that one interpretation is correct and the other erroneous; instead, it underscores the variability in applying different emotional criteria, both falling under the umbrella of negative emotions.

Upon deeper analysis, the discrepancy in interpretation may not solely be a flaw in ChatGPT4's functionality but could also stem from inadequate prompt design. As we incorporate more complex prompt word guides, it becomes increasingly challenging to encompass the nuanced emotional contexts with a limited set of instructions. This situation opens up avenues for future improvements: if strict adherence to the dataset's guidelines is not mandatory, enhancing the model based on broader prompt settings (such as more descriptive cues about people in images) could be a viable strategy. However, reliance on dataset labels for evaluation may be less effective, potentially necessitating a more comprehensive manual review process. Conversely, if absolute fidelity to the dataset's guidelines is essential, more than employing a few general prompt settings may be required. Instead, a more structured and supervised fine-tuning of the model may be necessary to ensure accuracy and adherence to specific emotional classifications.

VI. Conclusion and Future Work

In this paper, we studied the problem of emotion recognition, classification, and prediction. Emotion recognition situations have become increasingly important due to the role played by emotions in key personnel in how they perform their duties, which, particularly for safety-critical tasks, can mean jeopardizing or saving lives. To determine emotional competence in discharging duties, we group different emotion states into positive (i.e., competent) or negative (i.e., incompetent). We have developed a stochastic model to understand emotional dynamics and the evolution of emotional states. Our model is

able to represent the influences on how individuals experience and sustain their emotions over time and offers a structured way to investigate the intricacies of human emotions. By assigning numerical values to emotional decay rates, it is possible to quantify how quickly individuals return to a baseline emotional state after experiencing fluctuations. This approach allows for the observation and comparison of emotional stability across different individuals over periods.

A key parameter in this model is the emotion stability factor λ. Individuals with lower emotion stability factor values tend to have more stable emotional states. They can maintain a positive or neutral emotional condition for a longer duration, which suggests resilience to rapid shifts in mood due to external events or internal thoughts. Conversely, those with higher emotion stability factor values are more susceptible to emotional swings. Such sensitivity might be due to various factors, including external stimuli like social interactions or internal changes such as hormonal shifts.

In summary, using mathematical models to analyze and interpret the evolution of emotions based on emotion stability factor values provides a scientific basis for customized emotional health strategies. This approach enhances our understanding of emotional dynamics and supports the development of more effective psychological treatments and wellness programs tailored to individual needs. Such methodologies could lead to significant advancements in mental health practices, ultimately improving the quality of life for various populations.

Additionally, through experiments, we delve into the zero-shot capabilities of ChatGPT4 for image-based sentiment reasoning and judgment, comparing results with ResNet-50 and ViT models. Results show that while ChatGPT4's predictive power holds up well against the other two models, there is still much room for improvement, primarily by integrating mental health analysis and humanistic inputs. The main challenges identified include unstable predictions and inaccurate reasoning. Our results highlight the inherent difficulty of tasks such as mental health analysis and sentiment reasoning for image conversations, which remain daunting tasks for ChatGPT. However, we believe that further progress can be made to enhance the performance of ChatGPT4 through improved prompt engineering and more selective integration of contextual examples. Such enhancements are critical for their potential applications in real-world mental health settings and other related fields, where nuanced sentiment understanding is crucial.

In future work, we may fine-tune the ResNet-50 and ViT models to improve their generalization capabilities. For ResNet-50, this may involve using advanced regularization techniques, data augmentation, and carefully tuning hyperparameters to stabilize the training process. For ViT, increasing the number of training epochs, using transfer learning from larger datasets, and employing data augmentation may improve its performance. In addition, an ensemble approach that combines the predictions of ResNet-50, ViT, and GPT-4 may help. The ensemble can use ResNet-50 for detailed local feature extraction, ViT to capture global context, and GPT-4 to integrate external context and provide a multimodal perspective. By weighting the predictions according to each model's respective strengths, the ensemble can produce a more comprehensive and accurate emotion classification system.

References

[1] C. H. C. Leung and Z. Xu, "Emotional recognition and classification using large language models", in *The Ninth International Conference on Neuroscience and Cognitive Brain Information, BRAININFO*, ThinkMind, 2024, pp. 4–10.

[2] C. H. C. Leung, J. J. Deng, and Y. Li, "Enhanced human-machine interactive learning for multimodal emotion recognition in dialogue system", in *Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence*, 2022, pp. 1–7.

[3] J. J. Deng and C. H. C. Leung, "Towards learning a joint representation from transformer in multimodal emotion recognition", in *Brain Informatics: 14th International Conference, BI 2021, Virtual Event, September 17–19, 2021, Proceedings 14*, Springer, 2021, pp. 179–188.

[4] J. J. Deng, C. H. C. Leung, and Y. Li, "Multimodal emotion recognition using transfer learning on audio and text data", in *Computational Science and Its Applications–ICCSA 2021: 21st International Conference, Cagliari, Italy, September 13–16, 2021, Proceedings, Part III 21*, Springer, 2021, pp. 552–563.

[5] J. J. Deng and C. H. C. Leung, "Deep convolutional and recurrent neural networks for emotion recognition from human behaviors", in *Computational Science and Its Applications–ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20*, Springer, 2020, pp. 550–561.

[6] J. J. Deng and C. H. C. Leung, "Dynamic time warping for music retrieval using time series modeling of musical emotions", *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 137–151, 2015.

[7] J. J. Deng, C. H. C. Leung, A. Milani, and L. Chen, "Emotional states associated with music: Classification, prediction of changes, and consideration in recommendation", *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 5, no. 1, pp. 1–36, 2015.

[8] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, "Natural language processing applied to mental illness detection: A narrative review", *NPJ Digital Medicine*, vol. 5, no. 1, p. 46, 2022.

[9] D. Ciraolo *et al.*, "Emotional artificial intelligence enabled facial expression recognition for tele-rehabilitation: A preliminary study", in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2023, pp. 1–6.

[10] B. Mann *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020.

[11] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.

[12] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis", *arXiv preprint arXiv:2304.03347*, 2023.

[13] W. Zhao *et al.*, "Is chatgpt equipped with emotional dialogue capabilities?", *arXiv preprint arXiv:2304.09582*, 2023.

[14] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning", *IEEE Access*, vol. 11, pp. 14 742–14 751, 2023.

[15] P. Ekman, *Facial expressions of emotion: New findings, new questions*, 1992.

[16] P. Robert, *Emotion: Theory, research, and experience. vol. 1: Theories of emotion*, 1980.

[17] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context", in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1667–1675.

[18] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.

[19] Z. Lian, Y. Li, J.-H. Tao, J. Huang, and M.-Y. Niu, "Expression analysis based on face regions in real-world conditions", *International Journal of Automation and Computing*, vol. 17, pp. 96–107, 2020.

[20] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks", *Biomedical Signal Processing and Control*, vol. 59, p. 101 894, 2020.

[21] G. Trends and Forecasts, "Emotion detection and recognition market size and share analysis", 2024, [Online]. Available: https://www.mordorintelligence.com/industry-reports/emotion-detection-and-recognition-edr-market.

[22] N. Segal, *Facial Expressions Training Data*, https://www.kaggle.com/datasets/noamsegal/affectnet-training-data, 2022.

[23] ananthu017, *Emotion Dection*, https://www.kaggle.com/datasets/ananthu017/emotion-detection-fer, 2020.

[24] S. Vaidya, *Natural Human Face Images for Emotion Recognition*, https://www.kaggle.com/datasets/sudarshanvaidya/random-images-for-face-emotion-recognition, 2020.

[25] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild", *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[26] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].