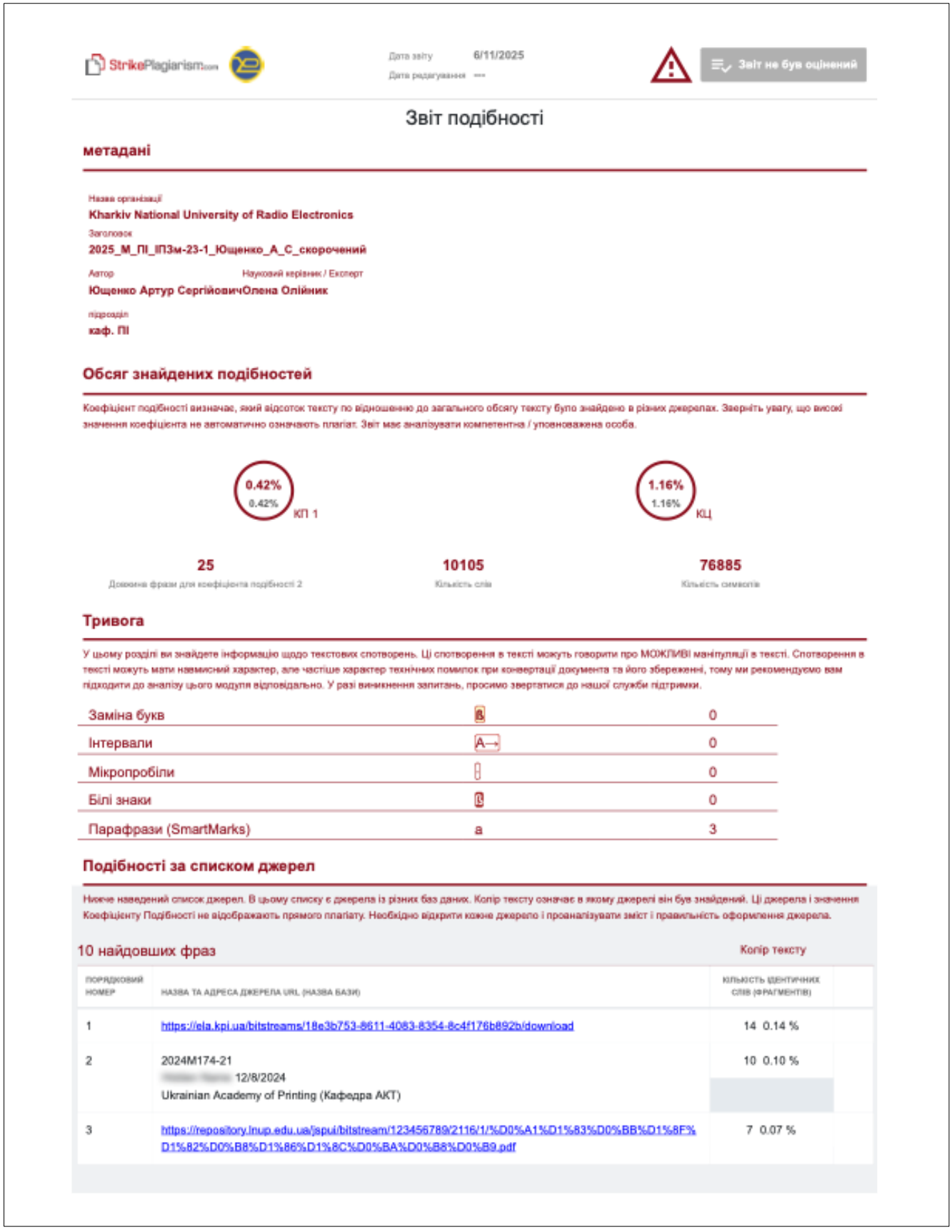
ДОДАТОК А

Звіт результатів перевірки на унікальність тексту в базі ХНУРЕ



StrikePlagiarism.com

Дата звіту 6/11/2025
Дата редагування ---

Звіт не був оцінений

Звіт подібності

метадані

Назва організації
Kharkiv National University of Radio Electronics
Заголовок
2025_М_ПІ_ІПЗм-23-1_Ющенко_А_С_скорочений
Автор Науковий керівник / Експерт
Ющенко Артур СергійовичОлена Олійник
підрозділ
каф. ПІ

Обсяг знайдених подібностей

Коефіцієнт подібності визначає, який відсоток тексту по відношенню до загального обсягу тексту було знайдено в різних джерелах. Зверніть увагу, що високі значення коефіцієнта не автоматично означають плагіат. Звіт має аналізувати компетентна / уповноважена особа.

0.42% 0.42% КП 1

1.16% 1.16% КЦ

25	10105	76885
Довжина фрази для коефіцієнта подібності 2	Кількість слів	Кількість символів

Тривога

У цьому розділі ви знайдете інформацію щодо текстових спотворень. Ці спотворення в тексті можуть говорити про МОЖЛИВІ маніпуляції в тексті. Спотворення в тексті можуть мати навмисний характер, але частіше характер технічних помилок при конвертації документа та його збереженні, тому ми рекомендуємо вам підходити до аналізу цього модуля відповідально. У разі виникнення запитань, просимо звертатися до нашої служби підтримки.

Заміна букв	a	0
Інтервали	A→	0
Мікропробіли		0
Білі знаки		0
Парафрази (SmartMarks)	a	3

Подібності за списком джерел

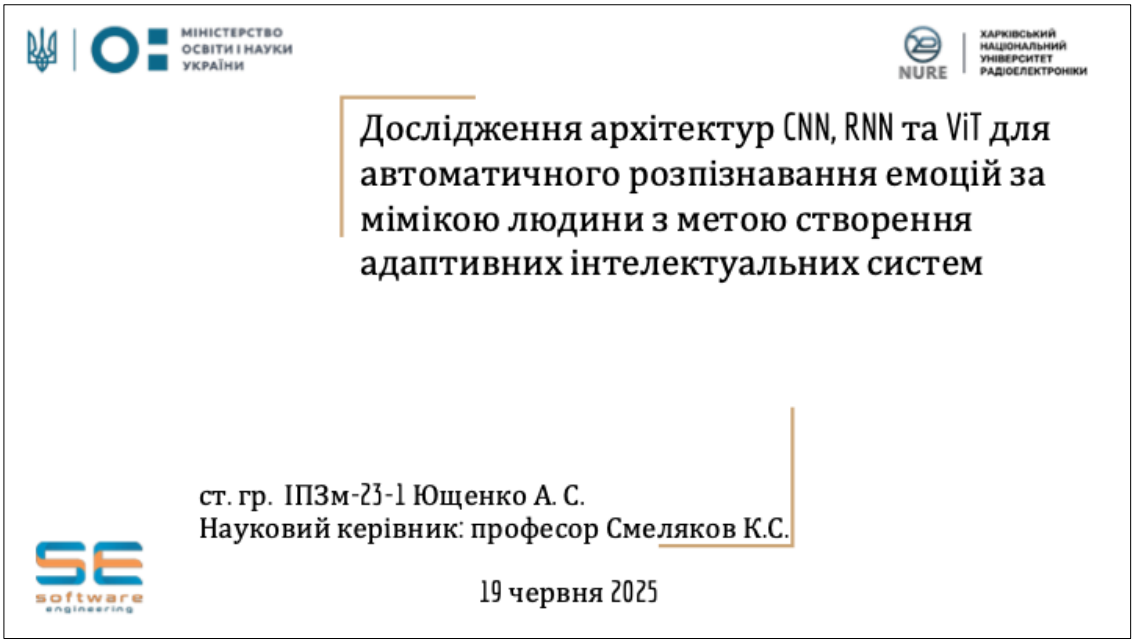
Нижче наведений список джерел. В цьому списку є джерела із різних баз даних. Колір тексту означає в якому джерелі він був знайдений. Ці джерела і значення Коефіцієнту Подібності не відображають прямого плагіату. Необхідно відкрити кожне джерело і проаналізувати зміст і правильність оформлення джерела.

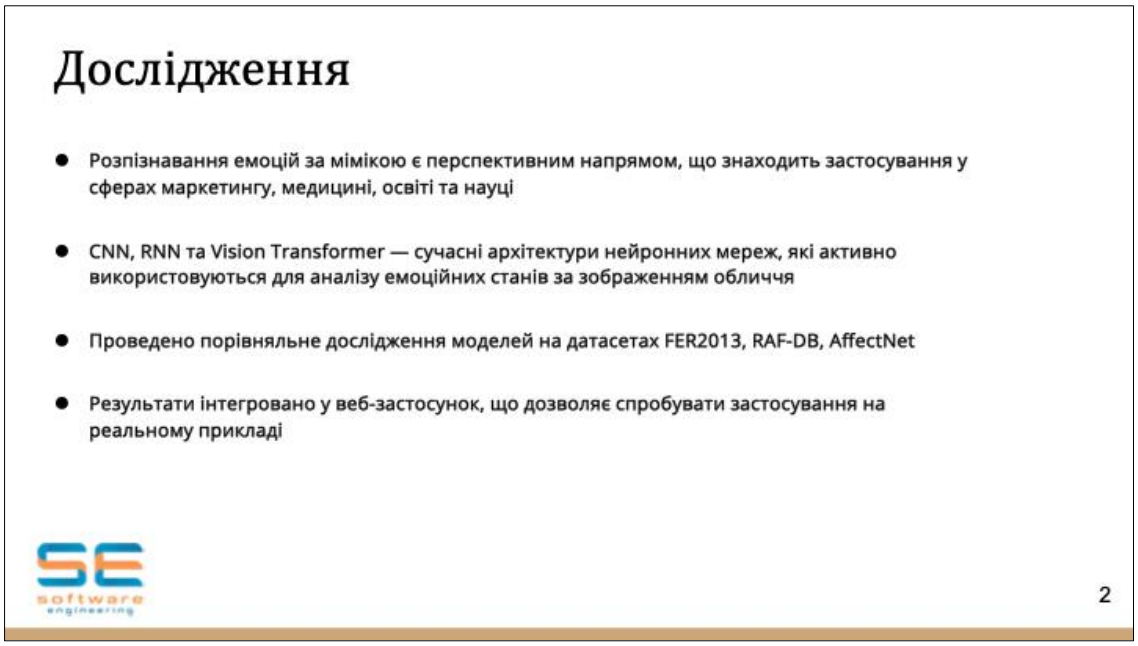
10 найдовших фраз — Колір тексту

ПОРЯДКОВИЙ НОМЕР	НАЗВА ТА АДРЕСА ДЖЕРЕЛА URL (НАЗВА БАЗИ)	КІЛЬКІСТЬ ІДЕНТИЧНИХ СЛІВ (ФРАГМЕНТІВ)
1	https://ela.kpi.ua/bitstreams/18e3b753-8611-4083-8354-8c4f176b892b/download	14 0.14 %
2	2024M174-21 [redacted] 12/8/2024 Ukrainian Academy of Printing (Кафедра АКТ)	10 0.10 %
3	https://repository.lnup.edu.ua/jspui/bitstream/123456789/2116/1/%D0%A1%D1%83%D0%BB%D1%8F%D1%82%D0%B8%D1%86%D1%8C%D0%BA%D0%B8%D0%B9.pdf	7 0.07 %

ДОДАТОК Б

Слайди презентації





Огляд літератури (аналогів)

Основою дослідження стали сучасні праці з аналізу міміки для розпізнавання емоцій:

- Facial Emotion Detection Using Deep Learning
- Recognition of Facial Expressions under Varying Conditions Using Dual-Feature Fusion
- Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions

Також використовувалась технічна документація: TensorFlow, Flask

Постановка задачі

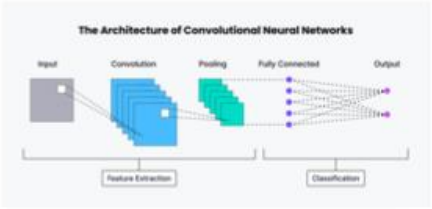
- порівняти ефективність архітектур нейронних мереж за ключовими метриками
- визначити переваги та недоліки на основі аналізу результатів
- розробити веб-застосунок для застосування у реальному житті
- сформувати рекомендації щодо вибору архітектури для роспізнавання емоцій

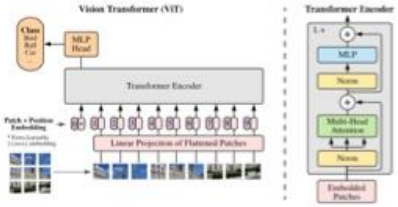
Методи дослідження

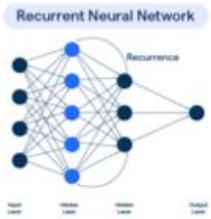
У дослідженні використано теоретичний аналіз та експериментальний підхід для практичної частини. В теоретичному аналіз розглянуто архітектури, їх особливості та архітектуру майбутнього веб-додатку. В практичній частині навчано три архітектури CNN, RNN, ViT на чотирьох дата сетах FER2013, RAF-DB, не повна версія AffectNet, повна версія AffectNet. Проведено тестування та порівняльний аналіз, за такими метриками: точність, площа під кривою (AUC), показник F1, затримка, пропускна здатність. Був створений веб-додаток для тестування всіх моделей на реальному прикладі інтеграції.

5

Архітектури нейроних мереж

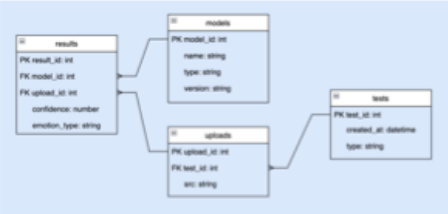






6

Архітектура та проектування веб-застосунку



7

Використані метрики

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)}{N}$$

$$AUC = \int_0^1 TPR(FPR^{-1}(t))\, dt$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$Latency = \frac{1}{N} \sum_{i=1}^{N} (t_{end}^{(i)} - t_{start}^{(i)})$$

$$Throughput = \frac{N}{\sum_{i=1}^{N} (t_{end}^{(i)} - t_{start}^{(i)})}$$

8

Використані дата сети

FER-2013
- 32 398 зображень
- 48 на 48 пікселів
- Чорнобілі

RAF-DB
- 15 300 зображень
- 100 на 100 пікселів
- Кольорові

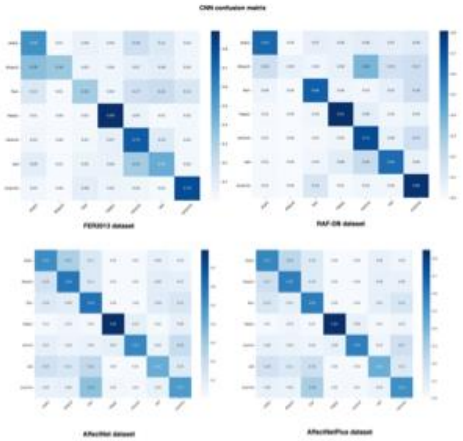
AffectNet неповний
- 14 600 зображень
- Розміри різні
- Кольорові

AffectNet повний
- 427 тисяч зображень
- Розміри різні
- Кольорові

9

Результати експерименту CNN архітектури

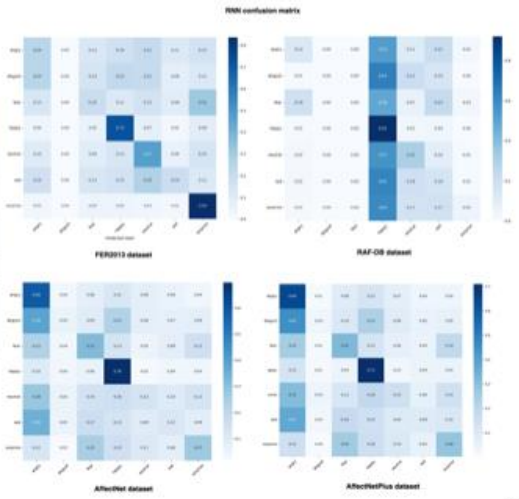
Метрика	FER2013	RAF-DB	Неповний AffectNet	Повний AffectNet
Accuracy (%)	64.47	70.53	54.17	56.73
AUC	0.9232	0.9307	0.8744	0.8862
F1-score	0.6343	0.7026	0.5399	0.5602
Latency (ms)	1.23	1.13	1.22	1.23
Throughput (images/s)	811.22	885.01	819.47	817.63



10

Результати експерименту RNN архітектури

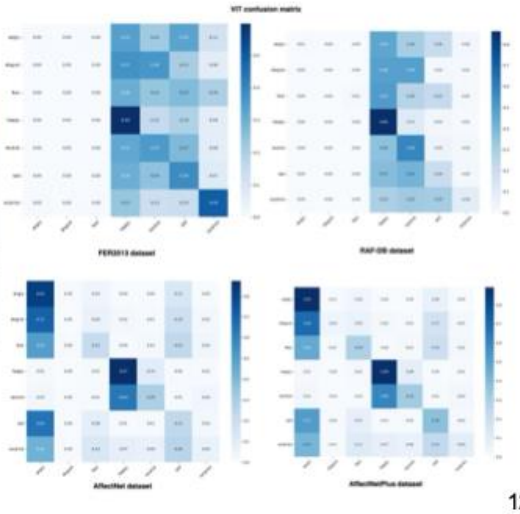
Метрика	FER2013	RAF-DB	Неповний AffectNet	Повний AffectNet
Accuracy (%)	45.01	46.48	33.73	35.93
AUC	0.8178	0.8203	0.7296	0.7412
F1-score	0.4225	0.3808	0.2961	0.3105
Latency (ms)	4.11	4.24	4.09	4.01
Throughput (images/s)	243.25	235.69	244.86	249.24



11

Результати експерименту ViT архітектури

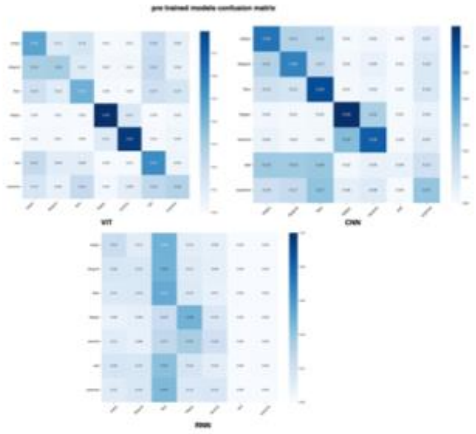
Метрика	FER2013	RAF-DB	Неповний AffectNet	Повний AffectNet
Accuracy (%)	31.60	50.03	19.90	23.51
AUC	0.7156	0.8394	0.5762	0.5803
F1-score	0.2475	0.4346	0.1509	0.1117
Latency (ms)	2.58	2.94	2.90	2.91
Throughput (images/s)	387.89	340.44	344.82	343.5



12

Результати предтренованих моделей

Метрика	ViT	CNN	RNN
Accuracy (%)	52.38	34.99	22.56
AUC	0.8724	0.7627	0.6275
F1-score	0.5162	0.3310	0.1825
Latency (ms)	34.70	11.02	8.72
Throughput (images/s)	342.1	191.69	114.67



13

Аналіз отриманих результатів

- CNN — оптимальний вибір, найшвидші, підходять для задач з обмеженими ресурсами
- ViT — доцільно використовувати тільки на дуже великих датасетах або використовувати предтреновані моделі
- RNN — не найкращий вибір для розпізнавання статичних зображень

14

Публікація результатів

Доступна за посиланням - https://ieeexplore.ieee.org/document/11016864

14

Висновки

- Проведено аналіз популярних архітектур нейронних мереж для розпізнавання емоцій CNN, RNN, ViT
- Проведено дослідження всіх трьох архітектур на чотирьох різних дата сетах
- Спроектовано та розроблено веб-застосунок для реального прикладу застосування всіх пятнадцяти навчених моделей
- На основі експерементів були сформовані рекомендації щодо вибору
- Була опублікована та презентована наукова стаття на конференції eStream 2025

16

Підсумки

Реалістичність та користь:
В результаті експерементів були проаналізовані реально важливі метрики при виборі архітектури для розпізнавання емоцій. Були використані різні дата сети які в своїх наборах мають різноманітні зображення.

Для майбутніх досліджень:
Спробувати використати гібридний підхід архітектур, наприклад поєднання ViT та CNN або використати мультиканальний підхід якщо є можливість брати данні не тільки з зображень.

ДОДАТОК В

Апробація результатів роботи

Evaluating CNN, RNN, and Vision Transformer for Emotion Recognition: Strengths and Weaknesses

Artur Yushchenko
Department of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
artur.iushchenko@nure.ua

Kirill Smelyakov
Department of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
kyrylo.smelyakov@nure.ua

Anastasiya Chupryna
Department of Software Engineering
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
anastasiya.chupryna@nure.ua

***Abstract*—This paper explores three prominent deep learning architectures — Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Vision Transformers (ViT) — for emotion recognition, examining their potential strengths and weaknesses under various conditions. It discusses how each approach may capture critical spatial, temporal, or global features in emotional data, highlighting differences in feature extraction, representational capacity, and scalability. Additionally, new solutions are proposed to enhance accuracy and adaptability, integrating design principles that address recognized challenges in real-world implementations. Novel insights are offered on aligning model selection with specific application demands, such as the nature of input signals, available computational resources, and desired real-time performance. While the comparative analysis remains broad to accommodate diverse use cases, it underscores the importance of carefully balancing accuracy and efficiency. Conclusions drawn from the investigation include recommendations on when each architecture may be most advantageous, providing a flexible framework for researchers and practitioners to navigate the trade-offs. These findings have implications for developing adaptive emotion recognition systems that leverage state-of-the-art deep learning techniques across multiple contexts.**

***Index Terms*—emotion recognition, deep learning, convolutional neural networks, recurrent neural networks, vision transformers, facial expression analysis**

I. Introduction

Emotion recognition has emerged as a pivotal research area, significantly benefiting from recent advances in deep learning. Modern algorithms routinely outperform earlier hand-crafted approaches by learning expressive features directly from data [1], [2]. Such developments have expanded applications across various domains, including human-machine interaction, mental health monitoring, adaptive user interfaces, and surveillance, by accurately interpreting human effects from faces, voices, and other modalities.

The evolution of deep learning has introduced several prominent architectures—namely Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and, more recently, Vision Transformers (ViTs). Each architecture uniquely captures different dimensions of emotional data, addressing spatial, temporal, and global feature interactions inherent in emotional expressions. In particular, CNN-based schemes have been explored with various preprocessing strategies for facial localization before classifying emotional states [3], while ViT-based models have been proposed to learn attention from facial cues under diverse conditions effectively [4]. Other recent methods integrate CNN features with recurrent modules enhanced by attention mechanisms to detect and classify emotional expressions robustly [5].

Despite these advancements, comprehensive evaluations comparing these architectures under uniform experimental conditions remain limited. Previous studies have predominantly analyzed individual models independently [3]–[5], leaving a critical gap in understanding their comparative performance and suitability for various real-world scenarios. This gap motivates our research, which systematically assesses the strengths and limitations of CNNs, RNNs, and ViTs in emotion recognition.

The following objectives guide our study:

- **Literature Context and Motivation:** We provide an in-depth review of existing studies, highlighting recent advances and persistent limitations within current emotion recognition methodologies. This comprehensive discussion emphasizes the necessity for integrative analyses that address trade-offs among model complexity, data requirements, real-time applicability, data augmentation techniques, and computational efficiency.
- **Comparative Analysis:** Through rigorous evaluations on standard emotion recognition benchmarks, we compare CNNs, RNNs, and ViTs to determine the architecture that best balances accuracy, computational demands, and real-world applicability. This comparative approach elucidates the relative merits and constraints inherent to each model.
- **Research Contribution:** Building upon established literature, our research provides straightforward, actionable guidelines for selecting optimal deep-learning architectures tailored to specific application demands and resource constraints.

In summary, this work bridges the existing research gap through a detailed comparative analysis of prevalent deep learning architectures, offering valuable insights and practical considerations for researchers and practitioners engaged in

emotion recognition tasks.

II. METHODS AND MATERIALS

This section describes the overall methodology, including the selection and configuration of deep learning models, the experimental setup, and the evaluation protocols used in our study.

A. Model Overview and Rationale

We compare three prominent deep-learning architectures for emotion recognition: CNNs, RNNs, and ViTs. Each model was chosen based on its unique capability to capture different aspects of the input data:

1) Convolutional Neural Networks: CNNs have been the workhorse of image-based emotion recognition (primarily facial expression recognition) due to their powerful feature learning capabilities [6]. Deep CNN models can automatically learn relevant facial features (such as shapes of eyes or mouth) from large image datasets, achieving high accuracy in classifying expressions like happiness, anger, or sadness. For instance, a deep CNN was the winning approach in the FER2013 facial expression recognition challenge on Kaggle, demonstrating the effectiveness of CNNs on that dataset [7].

Fig. 1 illustrates a typical CNN-based pipeline for facial emotion recognition.

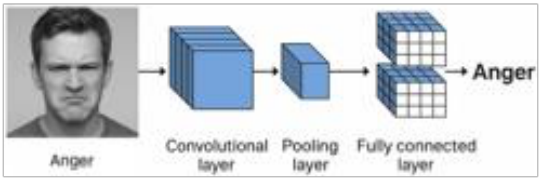


Fig. 1. Typical CNN pipeline for facial emotion prediction.

The input image is passed through convolutional layers that extract spatial features, followed by pooling layers that reduce dimensionality while retaining key information. The resulting feature maps are flattened and processed by fully connected layers to perform the final emotion classification.

Researchers have continued to improve CNN architectures for emotion recognition. R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata propose a novel deep CNN structure that outperforms prior models on benchmarks like RAF-DB (Real-World Affective Faces Database) [8]. Their improved network achieved higher accuracy on RAF-DB's 30,000 in-the-wild facial images by introducing enhanced convolutional blocks [6]. Similarly, other studies have shown that leveraging state-of-the-art CNN backbones (e.g., ResNet, EfficientNet) or fine-tuning pre-trained CNNs can significantly boost facial emotion recognition performance [2].

2) Recurrent Neural Networks: RNNs are widely used for their strength in modeling temporal dependencies, essential for capturing the dynamic changes in facial expressions over time. As shown in Fig. 2, RNN-based models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) receive sequential input features—typically extracted by a CNN—and process them through recurrent layers to capture temporal dynamics across frames.

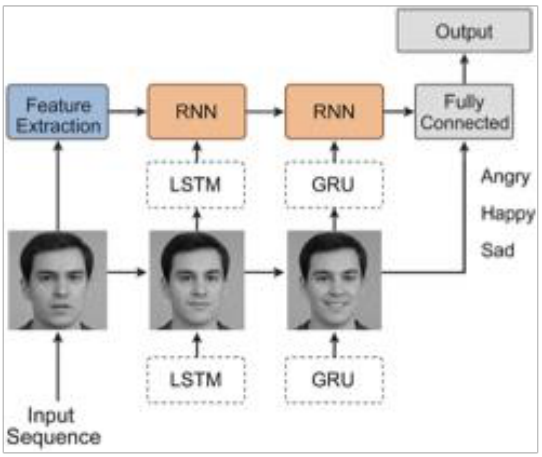


Fig. 2. Typical RNN architecture using LSTM and GRU layers for temporal modeling in facial emotion prediction.

We implement advanced RNN variants (LSTM/GRU) with extensive hyperparameter tuning, including hidden state size, number of recurrent layers, and dropout rates, to mitigate issues like vanishing gradients. These enhancements aim to optimize the network's ability to model the temporal evolution of expressions in video sequences.

In video-based emotion recognition, CNNs and RNNs (typically LSTMs) are frequently combined to leverage spatial and temporal features. Initially, CNNs extract frame-level facial features, which are subsequently fed into RNNs designed to learn the evolution of expressions across the sequence. M. Ye, H. Qian, and G. Liu introduced a CNN-LSTM framework equipped with a two-layer attention mechanism for facial expression recognition, achieving improved accuracy by prioritizing key frames within sequences [9]. Their hybrid model merges CNN-generated feature maps with an attention-enhanced LSTM that emphasizes salient facial changes, underlining the importance of detailed hyperparameter tuning (e.g., hidden state size, number of recurrent layers, dropout rates) to optimize the performance of these models for sequence data.

3) Vision Transformers: ViTs have recently gained popularity in computer vision and are now being explored for emotion recognition tasks. Unlike CNNs, which rely on local receptive fields, ViTs utilize self-attention mechanisms to model global relationships within an image. This capability makes them particularly effective for capturing spatial dependencies across different facial regions, such as combinations of action units or symmetric patterns. Early studies integrating transformers into facial expression recognition found that transformers can reach competitive performance with sufficient data or pretraining, though CNNs initially had an edge on smaller datasets [10].

As illustrated in Fig. 3, the input image is split into patches, linearly projected into embeddings, and passed through a transformer encoder composed of stacked self-attention layers. A classification head then predicts the corresponding emotion class.

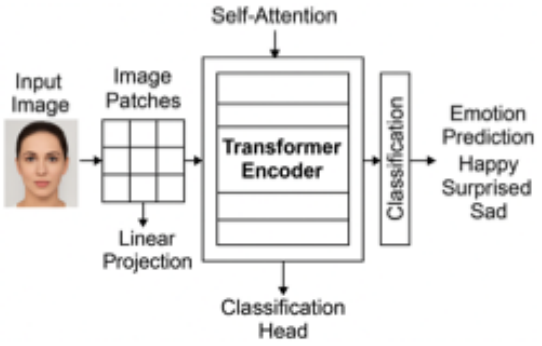


Fig. 3. Typical ViT architecture for facial emotion prediction using self-attention on image patches.

A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo showed that a fine-tuned ViT model (ViTFER) could perform on par with a CNN when trained on a large combined dataset of FER2013, AffectNet, and CK+ images [11]. In their work, they merged these datasets (creating an "AVFER" dataset of eight emotion classes) to mitigate data scarcity. They found the ViT achieved similar accuracy to a ResNet-18 baseline on the merged set. They also noted the ViT's ability to generalize across the merged dataset despite AffectNet's known label noise issues.

B. Datasets and Data Preprocessing in Affective Computing

A rigorous and reproducible experimental setup is essential for a fair comparison. Our setup includes the following components:

- **Hardware and Software Environment:** All experiments are conducted under standardized conditions on a MacBook Pro 16-inch (2021) equipped with the Apple M1 Pro chip (8-core CPU with six performance cores and two efficiency cores, 14-core GPU, 16-core Neural Engine, and 16GB unified memory). The software framework used is TensorFlow. This consistency ensures that any performance differences observed across models are attributable solely to architectural distinctions rather than hardware or software environment variations.
- **Data Acquisition and Preprocessing:** We utilize publicly available emotion recognition datasets, including FER2013 [12], AffectNet, and RAF-DB. The FER2013 dataset comprises approximately 30,000 grayscale images of facial expressions across seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutral expressions [13]. AffectNet contains over 41,000 facial images annotated for categorical emotions and valence-arousal dimensions, providing a comprehensive resource for affective computing research [14]. RAF-DB includes 29,672 facial images labeled with basic or compound expressions, offering diverse real-world facial expressions. Our data preprocessing pipeline involves standardizing pixel values through normalization, applying data augmentation techniques such as rotation, scaling, and flipping to enhance dataset diversity, and resizing all input images to match the required dimensions for each model. These preprocessing steps are essential for improving model generalization and performance.
- **Experimental Protocol:** Each model—CNN, RNN, and ViT—is trained and evaluated under identical data splits (same training, validation, and testing sets), uses the same batch size (32 images), and is run for a maximum of 50 epochs unless early stopping criteria are met.

C. Evaluation Metrics

To comprehensively evaluate model performance, we report the following metrics:

1) Accuracy: Accuracy measures the proportion of correctly classified samples among all evaluated samples:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i)}{N} \tag{1}$$

where N is the total number of samples, y_i is the actual label of sample i, $\hat{y}_i$ is the predicted label, and $\mathbb{1}(\cdot)$ is the indicator function (equal to 1 if its argument is valid, and zero otherwise).

2) Area Under the Curve (AUC): AUC assesses the trade-off between the true positive rate (TPR) and false positive rate (FPR) at varying decision thresholds:

$$\text{AUC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(t)\big)\, dt \tag{2}$$

In practice, the integral is often approximated numerically or via rank-based methods (e.g., the Mann–Whitney statistic).

3) F1-score: The F1-score is the harmonic mean of Precision and Recall, making it a balanced measure when class distributions are uneven:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

Precision (Prec) is the fraction of predicted positives that are genuinely positive, and Recall (Rec) is the fraction of actual positives that are correctly identified:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

4) Latency: Latency (in milliseconds) represents the average time taken by the model to process a single sample from input to output:

$$\text{Latency} = \frac{1}{N} \sum_{i=1}^{N} \left(t_{\text{end}}^{(i)} - t_{\text{start}}^{(i)} \right) \tag{4}$$

where $t_{\text{start}}^{(i)}$ is the time the i^{th} sample is fed into the model and $t_{\text{end}}^{(i)}$ is the time the i^{th} prediction is returned.

5) Throughput: Throughput (images/s) quantifies how many samples can be processed per unit time:

$$\text{Throughput} = \frac{N}{\sum_{i=1}^{N} \left(t_{\text{end}}^{(i)} - t_{\text{start}}^{(i)} \right)} \tag{5}$$

This metric is crucial for real-time or high-volume applications, where a higher throughput indicates more efficient batch processing.

D. Hyperparameter Tuning and Model Optimization

We not only adjusted general hyperparameters (e.g., number of layers, dropout rates, and learning-rate schedules) but also defined the exact model structures (input sizes, block configurations, and regularization strategies) to ensure reproducibility:

- **CNN Optimization:** We resized all input images to 48 × 48 (grayscale). The final CNN consisted of four sequential residual Squeeze-and-Excitation (SE) blocks, with filter sizes 32, 64, 128, and 256 (each using a 3 × 3 kernel). After each block, dropout rates incrementally increased from 0.2 up to 0.35. Following global average pooling, a fully connected (FC) layer with 512 units was applied before the output layer. An L2 regularization term was added to all convolution and dense layers [15].
- **RNN Optimization:** Again, using 48 × 48 grayscale images, we initially performed two convolutional blocks (32 and 64 filters) for spatial feature extraction. These feature maps were reshaped from 12 × 12 × 64 into sequences of length 12 (768 features each). Two stacked bidirectional LSTM layers of 128 units per direction were then employed to capture temporal dependencies, followed by a 256-unit FC layer with dropout [16].
- **ViT Optimization:** We divided each 48 × 48 image into patches of size 6 × 6, resulting in 64 patches. Four transformer blocks (each with 4 attention heads) were used, and the feed-forward sub-blocks had hidden dimensions of 128 followed by 64. A final MLP head of 128 units was applied before classification. As with the other models, dropout and L2 regularization were incorporated to mitigate overfitting [17].

III. Experimental Results

To thoroughly assess the performance of different deep learning architectures (CNN, RNN, and ViT) for emotion recognition, we conducted extensive experiments across three widely used datasets: FER2013, RAF-DB, and AffectNet. The training and evaluation processes were standardized to ensure fair and meaningful comparisons, and all experiments were executed using the Apple M1 Pro GPU.

This section comprehensively summarizes model performances, including accuracy, AUC, F1-score, latency, and throughput for each dataset. We present results separately for each architecture.

A. Convolutional Neural Networks

The CNN model employed a residual-based architecture to effectively capture spatial features, as detailed in Section II. Notably, the CNN achieved accuracies of 64.47% on FER2013, 70.53% on RAF-DB, and 54.17% on AffectNet, with corresponding AUC values of 0.9232, 0.9307, and 0.8744. The F1 scores were 0.5561, 0.7026, and 0.5399, respectively. Latency measurements were 14.17 ms (FER2013), 13.85 ms (RAF-DB), and 19.00 ms (AffectNet), while throughput reached 70.55 images/s (FER2013), 72.20 images/s (RAF-DB), and 52.62 images/s (AffectNet). A detailed class-wise analysis on AffectNet revealed a pronounced data imbalance, with "happy" emotions exhibiting strong recall in contrast to the negligible detection of "disgust," highlighting a key avenue for further refinement.

TABLE I
CNN Model Performance Across Datasets

Metric Name	FER2013	RAF-DB	AffectNet
Accuracy (%)	64.47	70.53	54.17
AUC	0.9232	0.9307	0.8744
F1-score	0.5561	0.7026	0.5399
Latency (ms)	14.17	13.85	19.00
Throughput (images/s)	70.55	72.20	52.62

B. Recurrent Neural Networks

An RNN architecture based on Long Short-Term Memory (LSTM) units was employed to capture temporal dependencies in sequential emotion data. The training methodology mirrored that used for the CNNs. As shown in Table II, the RNN attained accuracies of 45.01% on FER2013, 46.48% on RAF-DB, and 33.73% on AffectNet, with corresponding AUC scores of 0.8178, 0.8203, and 0.7296. The F1 scores were 0.42, 0.38, and 0.30, respectively. Latency values were measured at 23.50 ms, 22.80 ms, and 27.10 ms, while throughput reached 42.55, 43.85, and 36.90 images/s. These results underscore the higher computational cost of sequential data processing compared to CNNs.

TABLE II
RNN Model Performance Across Datasets

Metric Name	FER2013	RAF-DB	AffectNet
Accuracy (%)	45.01	46.48	33.73
AUC	0.8178	0.8203	0.7296
F1-score	0.42	0.38	0.30
Latency (ms)	23.50	22.80	27.10
Throughput (images/s)	42.55	43.85	36.90

C. Vision Transformers

The ViT architecture was examined for its ability to extract global image features. Table III presents its performance: accuracies of 31.60% on FER2013, 50.03% on RAF-DB, and 19.90% on AffectNet, alongside AUC values of 0.564, 0.8394, and 0.5762, respectively. The F1 scores were 0.25, 0.44, and 0.15. In terms of latency and throughput, ViT recorded

35.20 ms and 28.40 images/s (FER2013), 34.50 ms and 29.05 images/s (RAF-DB), and 41.80 ms and 23.95 images/s (AffectNet). These findings indicate that ViT performance suffers on smaller datasets or those with pronounced class imbalances.

TABLE III
ViT Model Performance Across Datasets

Metric Name	FER2013	RAF-DB	AffectNet
Accuracy (%)	31.60	50.03	19.90
AUC	0.564	0.8394	0.5762
F1-score	0.25	0.44	0.15
Latency (ms)	35.20	34.50	41.80
Throughput (images/s)	28.40	29.05	23.95

D. Discussion of Results

The results obtained from our extensive experimentation highlight distinct performance profiles for the three investigated architectures—CNNs, RNNs, and ViTs—across the FER2013, RAF-DB, and AffectNet datasets. CNNs consistently demonstrated the highest accuracy and AUC metrics, suggesting that their specialized convolutional filters and residual connections effectively capture spatial dependencies in static images. Furthermore, they showed the lowest latency and highest throughput, rendering them suitable for real-time or high-frequency emotion recognition tasks such as interactive facial expression applications and rapid psychological assessments. However, as evidenced by the class-wise evaluation on AffectNet, CNNs are not immune to the effects of substantial class imbalance, which can severely degrade their ability to classify underrepresented emotional classes correctly.

RNN-based models, leveraging Long Short-Term Memory (LSTM) cells, proved advantageous in scenarios where temporal patterns are critical, such as sequential frame analysis in video streams. Their ability to aggregate information over multiple time steps can be instrumental in capturing subtle inter-frame transitions of facial expressions. Nonetheless, the additional complexity associated with temporal modeling manifests in higher latency and reduced throughput compared to CNNs, raising significant concerns for latency-sensitive applications (e.g., real-time patient monitoring). This performance trade-off highlights the importance of carefully assessing the need for temporal modeling against available computational resources.

ViTs, while achieving state-of-the-art results in other computer vision tasks, exhibited lower accuracy and F1 scores in our emotion recognition experiments. Their reliance on global self-attention makes them powerful at modeling complex spatial correlations when trained on large-scale datasets. However, in the context of smaller emotion datasets or those with pronounced class imbalances, ViTs showed limited generalization capability and higher latency. This outcome underscores the necessity of larger training corpora or advanced data augmentation strategies to exploit ViTs for emotion recognition fully. Moreover, the computational overhead associated with transformer-based architectures further restricts their deployment in real-time, resource-constrained systems.

In summary, the choice of architecture should be tightly coupled with the specific requirements of the application domain. CNNs are an attractive default option, given their consistent robustness and computational efficiency. RNNs may be more suitable for long-term, sequential tasks, albeit at the cost of higher latency. ViTs, though promising, currently demand ample data and substantial computational resources, particularly when addressing highly imbalanced emotional categories.

IV. Discussion

This study provided a comprehensive comparative analysis of three major deep learning architectures—CNNs, RNNs, and ViTs—applied to emotion recognition across three benchmark datasets: FER2013, RAF-DB, and AffectNet. Our experiments were systematically conducted under uniform conditions to ensure fair comparisons and reliable results.

The empirical findings underscore several vital aspects relevant to research and practical applications. CNN architectures consistently outperformed RNNs and ViTs in accuracy, AUC scores, and computational efficiency across all datasets evaluated. This advantage is primarily attributed to their robust capability to extract spatial features from static images, efficiently capturing the nuanced facial expressions inherent in emotional data. Notably, CNNs delivered optimal real-time performance with latency and throughput measures far superior to those of RNNs and ViTs, making them particularly suitable for applications demanding real-time inference, such as interactive user interfaces and real-time mental health monitoring systems.

On the other hand, RNN architectures, particularly those employing LSTM units, demonstrated strength in capturing temporal dependencies present in sequential or time-series emotional data [18], [19]. Despite their comparatively lower throughput and higher latency—which limits their suitability for specific real-time applications—their inherent capability to handle dynamic sequences positions them advantageously for tasks involving continuous emotional monitoring, such as video analysis or longitudinal health tracking.

ViTs, while being at the forefront of current research in computer vision, showed relatively weaker performance in our emotion recognition benchmarks. The experiments revealed their accuracy and efficiency substantially declined, especially in datasets with limited size and pronounced class imbalances. These findings align with known limitations of transformer-based models, including their heavy reliance on large-scale datasets and extensive computational resources for training [20]. Consequently, their applicability appears restricted to scenarios where ample data and computing power are readily available [18], [21].

The class-wise performance analysis, evident in the AffectNet dataset, highlighted significant challenges posed by a severe class imbalance—a prevalent issue in real-world emotion recognition datasets. The stark disparity observed in

recall and precision across emotional categories (e.g., high recall in "happy" versus negligible performance in "disgust") underscores the urgent necessity of targeted data augmentation strategies, specialized training techniques (such as focal loss or balanced sampling), and potentially hybrid architectures to mitigate such imbalances effectively.

Given these insights, our study proposes several pathways for future improvements. For CNN-based models, we recommend further refinement through advanced augmentation techniques and adaptive learning strategies to address the class imbalance, thus enhancing their already strong baseline performance. For RNNs, exploring optimized model structures and hybrid temporal-spatial frameworks may help offset their computational inefficiencies, broadening their practical applicability. Lastly, to unlock the potential of ViTs for emotion recognition, future work must focus on tailored pretraining strategies, sophisticated regularization methods, and efficient architectures to manage their computational complexity better.

V. Conclusions

Ultimately, the decision to employ a particular deep learning architecture for emotion recognition tasks should be guided by a careful assessment of application-specific requirements such as the nature and volume of input data, available computational resources, and the criticality of real-time inference. Our findings show that CNNs consistently deliver robust accuracy and efficiency for static images, RNNs better capture temporal dynamics in sequential data (albeit at a higher computational cost), and ViTs offer strong global feature modeling when ample data and resources are available. This unified perspective gives practitioners and researchers actionable insights to select and optimize models tailored to their unique demands effectively.

In summary, we recommend CNNs for real-time, resource-constrained applications with predominantly static inputs, RNNs for tasks requiring detailed temporal analysis over continuous sequences, and ViTs for large-scale or highly varied datasets where global context is paramount. These recommendations can be refined through specialized data augmentation, balanced training strategies, and hybrid architectures to mitigate class imbalance. In conclusion, this research contributes significantly to the ongoing discourse in emotion recognition by elucidating the strengths and limitations of prevalent deep learning architectures and establishing clear guidelines for informed architecture selection, thereby paving the way for more robust, accurate, and efficient emotion recognition systems.

References

[1] R. R. Adyapady and B. Annappa, "A comprehensive review of facial expression recognition techniques," Multimedia Systems, vol. 29, no. 1, pp. 73–103, 2023.

[2] H. Ge, Z. Zhu, Y. Dai, B. Wang, X. Wu, "Facial expression recognition based on deep learning," Computer Methods and Programs in Biomedicine, vol. 215, Art. no. 106621, 2022.

[3] R. Dalal, M. Khari, P. Pandey, S. Jatana, V. Joshi, "Facial Emotion Recognition and Detection Using Convolutional Neural Networks," in IEEE Access, Feb. 2024, pp.1-8.

[4] A. Chowanda, Nadia, Diana, "Implementing Vision Transformer to Model Emotions Recognition from Facial Expressions," in IEEE Access, Sep. 2023, pp. 48-53.

[5] S. V. Kumar, G. Sunil, R. R. Hussein, S. M. Vidhya, S. M. Sundaram, "Attention based ConVnet-Recurrent Neural Network for Facial Recognition and Emotion Detection," in IEEE Access, Oct. 2024, pp. 1-5.

[6] R. A. Elsheikh, M. A. Mohamed, A. M. Abou-Taleb, and M. M. Ata, "Improved facial emotion recognition model based on a novel deep convolutional structure," Scientific Reports, vol. 14, no. 1, Art. no. 29050, Nov. 2024.

[7] J. R. Lee, L. Wang, and A. Wong, "EmotionNet Nano: An efficient deep convolutional neural network design for real-time facial expression recognition," Frontiers in Artificial Intelligence, vol. 3, Art. no. 609673, 2021.

[8] A. Sultana, S. K. Dey, and M. A. Rahman, "A deep CNN based Kaggle contest winning model to recognize real-time facial expression," in Proc. 5th Int. Conf. Electrical Engineering and Information Communication Technology (ICEEICT), 2021.

[9] M. Ye, H. Qian, and G. Liu, "CNN-LSTM facial expression recognition method fused with two-layer attention mechanism," Computational Intelligence and Neuroscience, vol. 2022, Art. no. 7450637, 2022.

[10] F. Xue, Q. Wang, Z. Tan, Z. Ma, G. Guo, "Vision Transformer with Attentive Pooling for Robust Facial Expression Recognition," IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 3244–3256, 2023.

[11] A. Chaudhari, C. Bhatt, A. Krishna, and P. L. Mazzeo, "ViTFER: Facial Emotion Recognition with Vision Transformers," Applied System Innovation, vol. 5, no. 4, Art. no. 80, 2022.

[12] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Transactions on Affective Computing, vol. 13, no. 3, pp. 1195–1215, Oct. 2018.

[13] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression," in Proc. IEEE Workshop on CVPR for Human Communicative Behavior Analysis, 2010.

[14] 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, Oct. 2017.

[15] A. Jain, "All about convolutions, kernels, features in CNN," Medium. Accessed: Mar. 28, 2025. [Online]. Available: https://medium.com/@abhishekjainindore24/all-about-convolutions-kernels-features-in-cnn-c656616390a1

[16] S. Nath, "Regularization Techniques: Preventing Overfitting in Deep Learning," Medium. Accessed: Mar. 28, 2025. [Online]. Available: https://medium.com/@sruthy.sn91/regularization-techniques-preventing-overfitting-in-deep-learning-2cddb7822217

[17] S. Chhabra, H. Venkateswara, and B. Li, "PatchSwap: A regularization technique for vision transformers," in Proc. British Machine Vision Conf. (BMVC), London, UK, Nov. 2022, pp. 1-12.

[18] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," Scientific Reports, vol. 14, no. 21392, Sep. 2024.

[19] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," Applied Sciences, vol. 13, no. 9, p. 5521, Apr. 2023.

[20] S. Hazmoune and F. Bougamouza, "Using transformers for multimodal emotion recognition: taxonomies and state of the art review," Engineering Applications of Artificial Intelligence, vol. 133, Art. no. 108339, 2024.

[21] H. Idrees, "Vision Transformer vs. CNN: A Comparison of Two Image Processing Giants," Medium. Accessed: Mar. 28, 2025. [Online]. Available: https://medium.com/@hassaanidrees7/vision-transformer-vs-cnn-a-comparison-of-two-image-processing-giants-d6c85296f34f

ДОДАТОК Г

Експертний висновок результатів перевірки кваліфікаційної роботи на відповідність оформлення вимогам ДСТУ 3008: 2015

1

Експертний висновок результатів перевірки кваліфікаційної роботи

студент (посада) | програмної інженерії (кафедра) | ІПЗм-23-1 (група)

Артур ЮЩЕНКО
(прізвище, ім'я, по батькові)

Зауваження

Пункт ДСТУ 3008-2015	Зміст пункту	Сторінка кваліфікаційної роботи
1	2	3
	7.1 Загальні положення	
7.1.25	Не дозволено розміщувати назву розділу, підрозділу, а також пункту й підпункту на останньому рядку сторінки.	25
	7.3 Нумерація сторінок звіту	
	7.5 Рисунки	
	7.6 Таблиці	
	7.7 Переліки	
	7.8 Примітки	
	7.9 Виноски	
	7.10 Формули та рівняння	
	7.11 Посилання	
	7.13 Список авторів	
	7.14 Скорочення та умовні познаки	
	7.15 Додатки	
Методичні вказівки до виконання кваліфікаційної роботи магістра… **ЗАТВЕРДЖЕНО** кафедрою ПІ протокол № 5 від 13.11.2023р. 3.2 Оформлення пояснювальної записки згідно з ДСТУ 3008:2015 Звіти у сфері науки і техніки. Структура та правила оформлювання. **Шаблон** затверджений засіданням кафедри №3 від 16.10.2023.	Рисунок повинен розміщуватися одразу після його згадування у тексті, або на наступній сторінці. Під рисунком повинен бути підпис із словом Рисунок, порядковим номером цього рисунку, через тире з великої літери – назва рисунку та **в круглих дужках вказується джерело з якого взятий цей рисунок, або то, що його виконано самостійно.**	за текстом

Експерт _____ (підпис) Вадим НЕЧВОЛОД (прізвище, ініціали)

Вкажіть джерела походження рисунків.

12.06.2025