جامعة أم القرى
UMM AL-QURA UNIVERSITY

Machine Learning Project: Stress detection using Face Analysis

غدير سامي نفادي - 44511798

منار تركي أحمد - 445006780

رتاج حسين الحازمي - 445006594

غلا شاهر البشري - 445006326

روان محمد قادر - 445013919

Group: 2

Supervisor: Dr. Afaf Al -Mehmadi

Abstract

This project explores the development of a ML module for automatic stress detection from facial expressions. The motivation comes from the growing need to help psychologists identify early signs of stress and anxiety. These signs are often subtle and harder to notice compared to basic emotions (anger or happiness). We started with a video-based. Aimed at capturing stress cues over time, but the results limited due to noisy labels, inconsistencies between frames, and computational challenges. As a result, we chose a more stable image-based approach using the FER-2013 dataset. We reorganized the dataset's 7-emotion labels into 2-meaningful categories: Stress (angry, disgust, fear) and No-Stress (happy, neutral). We trained a (CNN) on the reorganized dataset with data augmentation to improve generalization. The proposed CNN showed promising performance, achieving around 79% accuracy, an F1-score of about 0.73, and an AUC of roughly 0.88 on the test set. Results highlight the potential of using facial expressions to detect stress as a helpful tool for mental health applications. The report includes a literature review, gap analysis, comparison of methods, limitations, and suggestions for future multimodal systems that combine facial, audio, and physiological signals for more accurate stress assessment.

Keywords: stress detection, facial expression analysis, CNN, FER-2013, mental health, machine learning.

Introduction

In our daily lives, we all go through situations that are funny, sad, happy, angry, or make us feel stressed and anxious. But sometimes the situation worsens, and we need the intervention of a psychologist, either because of excessive stress, anxiety, sadness, or other reasons.

However, it is easy to notice anger and sadness, unlike anxiety, psychological pressure, and stress. For this reason, researchers began developing and training AI models capable of analyzing facial features and predicting a person's emotion happy, sad, angry, and so on. This has contributed greatly to helping and developing technologies. But the world is evolving faster, and people have become more exposed to anxiety and stress, and psychologists need support in diagnosing or noticing signs of anxiety and stress in patients. Mood or emotion detection systems do not help them significantly, so it is important to develop models and datasets that assist them in diagnosing and noticing signs of stress in visitors and patients. In this paper, we focus on designing a machine learning module that automatically detects stress versus non-stress from facial expressions. We initially experimented with a video-based dataset to capture temporal signs of stress, but the results were poor, so we moved to a more stable image-based approach using the FER-2013 dataset. We reorganize its seven emotion labels into two groups (Stress vs. NoStress), train and evaluate a convolutional neural network (CNN) on this data, and present a literature review, gap analysis, and detailed description of the proposed module as a potential support tool for psychologists and mental health applications.

Problem definition

There are many facial and emotion recognition systems used in various fields, but there are no recognition systems specialized in deep emotions, reactions, and feelings that can help psychologists improve the diagnosis of patients with anxiety and stress disorders, and detect and observe some signs early. However, the problem is that building such systems is very expensive and difficult, and the current existing systems can only predict very basic and obvious emotions such as anger and happiness. They are neither accurate nor specialized in analyzing emotions such as psychological pressure, anxiety, and stress.

Therefore, it is necessary to develop a specialized system for recognizing anxiety and stress emotions and the signals and indicators related to them, in order to support psychologists and improve the mental health of individuals in society.

In this project, we propose a stress detection module that classifies a person's face as stress or no stress. It uses a convolutional neural network (CNN) trained on a reorganized version of the FER-2013 dataset, where certain emotions are grouped into these two classes. The goal is to provide an automatic tool that can help highlight possible stress cases and support psychological assessment.

Literature Review

a. Current State of the Art:

The landscape of automated stress detection has shifted significantly towards Deep Learning, with Convolutional Neural Networks (CNNs) establishing themselves as the gold standard for extracting spatial features from facial imagery. While some approaches rely on complex physiological sensors or thermal imaging, the current trend in accessible telemedicine focuses on optical-based detection using standard cameras.

A key methodology in this domain is the "Emotive-Proxy Approach," where specific higharousal negative emotions (such as Anger, Fear, and Disgust) are aggregated to model a clinical

"Stress" state. This approach moves beyond multiclass emotion recognition, which

often suffers from ambiguity, to a more robust binary risk assessment. By leveraging CNN architectures trained on large-scale datasets like FER-2013, and enhancing them with Data

Augmentation techniques to simulate real-world variability (e.g., rotation and pose shifts), current systems can effectively identify visual biomarkers of stress in a non-invasive and computationally efficient manner.

b. Comparative Analysis:

1- compare and contrast

Over the past decade, facial emotion recognition and stress detection have become central topics in affective computing.

Researchers have moved away from traditional approaches that relied on handcrafted features or intrusive physiological sensors, toward deep learning models capable of automatically extracting meaningful patterns from facial data. Convolutional Neural Networks (CNNs), deeper architectures such as ResNet and VGG, and lightweight models like MobileNetV2 have all been explored. Despite their differences, these studies share a common ambition: to build systems that can operate reliably in real-world conditions, where lighting, pose, cultural variation, and noisy labels present constant challenges.

The reviewed studies highlight a wide spectrum of models. ResNet-18, used as a baseline, offered simplicity and speed but struggled with small datasets such as AFEW-VA because it lacked landmark integration or attention mechanisms. In contrast, EmoFAN represented a significant advance: by combining facial landmark estimation, attention derived from heatmaps, and multi-task learning, it simultaneously predicted landmarks, discrete emotions, and continuous valence/arousal. This integration produced state-of-the-art results and clearly outperformed ResNet-18 .

Other work experimented with CNNs enhanced by edge detection, where Kirsch filters added structural information to feature maps. This approach achieved 88.56% accuracy, surpassing R-CNN (79.34%) and FRR-CNN (70.63%), demonstrating that even modest architectural innovations can yield substantial gains. Customized CNNs with bypass layers and Haar Cascade detection also proved effective, reaching over 92% accuracy on CK+, while maintaining efficiency through techniques such as K-fold cross-validation and Adam optimization. Stress detection studies introduced further diversity. One group employed SVM classifiers with thermal imaging, segmenting the face into five regions and using average

temperature values as features. Despite a small sample size of 25 participants, accuracy reached 95.45%. Others turned to MobileNetV2 with transfer learning on FER2013, which outperformed ResNet50 in speed and efficiency, even though ResNet50 remained superior in fine grained feature extraction. Hybrid models combining residual networks with backtracking also emerged, preserving information during training and outperforming standard CNNs, particularly after oversampling corrected class imbalance.

Finally, VGG16 with fine-tuning delivered strong results, achieving nearly 90% accuracy in multi-class emotion recognition and over 92% in binary stress classification.

Performance varied across datasets and architectures. ResNet-18 consistently underperformed, while EmoFAN achieved CCC improvements of up to +0.20 after label cleaning, setting new benchmarks on AffectNet and SEWA. CNNs with edge detection demonstrated clear advantages over R-CNN and FRR-CNN, both in accuracy and training speed. Customized CNNs achieved 86.78% on FER13, 92.27% on CK+, and 91.58% on JAFFE, underscoring the importance of dataset characteristics .

In stress detection, SVM with thermal imaging excelled despite limited data, while MobileNetV2 with transfer learning proved more practical than ResNet50 for real-time use. Hybrid models improved test accuracy from 78% to 84% after oversampling, highlighting the impact of data balancing. VGG16 with fine-tuning stood out as a versatile option, performing strongly across both multi-class and binary tasks.

Despite encouraging results, several limitations recur across studies:

- Data constraints: Small or imbalanced datasets (CK+, JAFFE) hinder generalization, while larger ones (FER 2013, AffectNet) suffer from noisy labels and annotator disagreement.
- Cultural bias: Training data often reflects limited demographics, reducing global applicability.
- Environmental sensitivity: Lighting, pose, occlusion by glasses or hair all degrade performance.
- Computational demands: Deep models like ResNet50 deliver precision but are resource-intensive; lighter models trade accuracy for speed.
- Stress labeling gaps: Some systems inferred stress from negative emotions rather than using direct stress annotations.
- Emotion overlap: Models struggled to separate similar categories, such as fear vs. sadness or anger vs. disgust.

- Hardware limitations: Reliance on specialized devices (e.g., thermal cameras) restricts deployment outside controlled environments.

Conclusion:

Together, these ten studies illustrate both the promise and the challenges of deep learning for emotion recognition and stress detection. Each model offers distinct trade-offs: ResNet and VGG deliver strong feature extraction but demand heavy computation; MobileNetV2 balances speed and accuracy for real-time use; CNNs with edge detection introduce creative enhancements; SVM with thermal imaging provides a non-traditional yet effective alternative; and hybrid models offer middle-ground solutions. Looking ahead, researchers emphasize the need for larger, more diverse datasets, improved handling of noise and cultural bias, and generative approaches such as GANs or VAEs to balance data. Ultimately, the goal is to design models that combine accuracy, efficiency, and robustness, enabling practical applications in education, mental health monitoring, and security.

2- Gap Analysis and justification

There are many studies available in this field, but most of them are limited to basic emotions such as joy, sadness, anger, and fear without direct connection to clinical diagnoses of anxiety, stress, or psychological pressure, which are often assumed through reactions or self assessment, and this cannot be considered a reliable reference. Also, most models are trained on images, therefore not applicable in thermal contexts and weak in generalization because the data they are trained on is clean, organized, and well structured, while in real life human races and cultures differ and there are many factors that affect the models such as lighting, head movement, wearing a mask or glasses, or placing the hand on the face. Images also ignore important indicators such as sweating and heartbeats, which cannot be measured with expensive devices for every patient. Static images are good for strong and clear expressions that are unambiguous, but they do not capture the small and fast signs associated with stress and anxiety, therefore many images expressing anxiety are misclassified as another emotion such as sadness due to the absence of small time related signs.

This is one of the most important observations: training on images does not provide comprehensiveness and clarity for anxiety expressions so they can be captured and recognized. Also, the data is not always real, as many models are built on data taken from
standard emotion datasets which are often fabricated or very small, and many participants are students or images from the internet with acted expressions, meaning unnatural.
This makes the models far from the reality of the patient's condition. Moreover, many doctors do not pay much attention to how the model reached this decision or whether it is correct or wrong.

Importance:

We do not want a system that recognizes smiles and frowns, but rather a tool that helps the doctor understand anxiety and stress in a real way with approved evaluations and standards. Therefore, it is important to collect data closer to reality with different expressions and videos that capture signs of anxiety, with diverse participants from different nationalities, races, and ages, and with clear labels taken from approved diagnoses**.**

3- Tables / Graphs

Table. 1. Papers summary

Limitations	Strengths	Results	Model	Dataset	Title
Requires good landmark detection, label noise, cultural bias	Multi-task learning, attention, landmarks, strong generalization	SOTA: +0.11 to +0.22 CCC improvements	EmoFAN (multitask CNN), ResNet-18 baseline	AffectNet, AFEW-VA, SEWA	Estimation of Continuous Valence & Arousal in Naturalistic Conditions
Small dataset, thermalcamera dependency, sensitive to ROI accuracy	Non-invasive, simple features, very high accuracy, real-time friendly	95.45% accuracy	HOG + SVM (ROI), SVM classifier	Custom thermal dataset (25 subjects)	Automatic segmentation of Facial ROIs & Stress Detection Using ML
Sensitive to pose/occlusion, limited generalization	High robustness in complex backgrounds, fast training	88.56% accuracy, faster than R-CNN	7-layer CNN + Kirsch edge detection	FER-2013, LFW	Face Emotion Recognition Using CNN + Image Edge Computing
Imbalanced datasets, confusion between fear/sadness	Lightweight architecture, K-fold validation, strong performance	86.78%–92.27% accuracy	Custom CNN + Haar Cascade + Data Generator	FER13, CK+, JAFFE	Deep Learning for Facial Emotion Recognition Using Custom CNN
Dependent on FER-2013 only, subtle emotions difficult	Very fast, lightweight, high accuracy	Higher accuracy than ResNet50 & Sequential	ESCNN (MobileNetV2 + TL)	FER-2013	ESCNN for Stress Detection (MobileNetV2-based)
FER-2013 noise, hybrid more complex	Oversampling improves results, hybrid captures deeper features	CNN: 78% test – Hybrid: 84% test	CNN vs Hybrid ResNet + Backtracking	FER-2013	Stress Level Classification Using FER-2013
No stress-specific labels, pose/lighting issues	Diverse datasets, real time capable	89.6% multiclass, 92.1% binary	CNN (VGG16, VGG19, Inception-ResNet) + TL	KDEF, CK+, Net Images	Facial Expression Recognition System for Stress Detection (VGG16/19)
Sensitive to lighting, occlusion, multi-face issues	Very robust, diverse dataset, real-time performance	96% accuracy	CNN + VGG16 TL	CK, JAFFE, Internet images (~10k)	Mental Health Detection Using Facial Emotion Recognition
Small EEG dataset, subjective labeling	Extremely fast, reduced feature set, high accuracy	99.81% CNN facial accuracy; 87.25% EEG LSTM	CNN (landmarks), LSTM (EEG)	Custom (55 students) + EEG subset	Real-Time Emotion Recognition (10 FACS markers)
High computation needs, data-hungry, micro-expressions difficult	DL highest accuracy, strong for video (LSTM)	SVM: 8%CNN: 99.32% hybrid: 98.72%	SVM/PCA, CNN, CNN-LSTM hybrid	CK+, Bosphorus, SMIC	Conventional ML vs Deep Learning in FER

Detailed ML Module (Code and Dataset)

We initially experimented with a video-based approach using the RAVDESS (Ryerson AudioVisual Database of Emotional Speech and Song) dataset and a MobileNetV2 model to capture temporal patterns of stress from facial expressions over time. However, the evaluation results were not very satisfactory, and the training process was computationally expensive and timeconsuming, we expect that allocating significantly more training time and tuning might improve performance, but this was not practical within our project constraints. Therefore, we shifted to a more efficient image-based setup using the FER-2013 dataset with a CNN model, where the evaluation results were noticeably better and the training process was more stable and manageable.

Phase I: Preliminary Experiment (Video-Based):

- **Dataset:**

  For this work, we rely on the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. It is a validated multimodal database widely used for training and testing emotion recognition models in both audio and video formats.

  It contains high-quality video recordings of 24 professional actors (12 male and 12 female) delivering statements with varying emotional intensity. Overall, the dataset includes eight distinct emotions, such as happy, sad, angry, and fearful, which serve as the foundation for classifying stress versus non stress states in our system.

- **Model:**

  We rely on the MobileNetV2 architecture. It is a lightweight Convolutional Neural Network (CNN) pre trained on the ImageNet dataset, widely used for efficient image analysis on resource-constrained devices.

  The model utilizes a Transfer Learning approach, where the pre trained network serves as a powerful feature extractor for facial expressions. We modified the final classification layers to output a binary probability, allowing the system to distinguish between Stress and Non Stress states with high efficiency.

```
Confusion matrix:
[[652 188]
 [382 578]]

Classification report:
              precision    recall  f1-score   support

  non-stress       0.63      0.78      0.70       840
      stress       0.75      0.60      0.67       960

    accuracy                           0.68      1800
   macro avg       0.69      0.69      0.68      1800
weighted avg       0.70      0.68      0.68      1800
```

Fig. 1. MobileNetV2 Model Evaluation

Phase II: Proposed System (Image-Based):

Dataset:

1. Dataset Overview

For this work, we rely on the FER-2013 (Facial Expression Recognition 2013) dataset. It is a public dataset created for the ICML 2013 Facial Expression Recognition Challenge. It contains faces showing different emotions, and it is widely used for training and testing emotion recognition models.

Overall, FER-2013 has around 35,887 face images, each labeled with one of seven basic emotions.

2. Data Source and Collection

The images in FER-2013 were collected from the internet, not from a lab. The creators used the Google Image Search API and searched for many different emotion-related words (for example “happy woman”, “angry man”, “surprised face”, etc.).

The downloaded images were processed in the following way:

- A face detector (based on OpenCV) was used to find and crop faces from the original images.
- The faces were then resized to a fixed size and converted to grayscale.
- Human annotators checked the faces and assigned an emotion label to each image (angry, happy, etc.). Some examples that did not show a clear face or emotion were removed.

Because the data comes from the real world (different cameras, lighting, angles, ages, etc.), it is considered a "in the wild" dataset and is more challenging than clean laboratory images.

3. Image Format and Size

Every image in FER-2013 has the same format:

- Resolution: 48 × 48 pixels.
- Color: grayscale (only one channel, no RGB).

This means each image is a small black-and-white picture of a face. The low resolution makes the task harder, because fine details (like wrinkles or small changes around the eyes) are not very clear.

4. Emotion Classes

Each image is labeled with one of seven basic emotion classes:

Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral.

These emotions come from the classic basic-emotion theory (Ekman's emotions), and they are commonly used in facial emotion research.

The dataset is not perfectly balanced. For example, there are usually more "happy" and "neutral" images than "disgust" images. Some research papers show that "disgust" has much fewer samples compared to other classes.

5. Dataset Structure and Splits

Originally, FER-2013 was released in CSV format with three main splits:

- Training set: about 28,709 images.
- Public test set (validation): about 3,589 images.
- Private test set: about 3,589 images.

Each entry in the CSV file includes:

- A "pixels" field: the 48×48 grayscale image flattened as a long list of pixel values.
- An "emotion" field: an integer from 0 to 6 representing the class (angry, disgust, fear, happy, sad, surprise, neutral).

On Kaggle, different versions of FER-2013 may be provided in a more convenient folder structure (for example, images grouped in folders by emotion and split into train and test). However, they all come from the same original FER-2013 CSV data.

6. Class Balance and Dataset Challenges

Because the dataset was collected from the internet, there are several challenges and limitations:

- Imbalanced classes: Some emotions (like “happy” and “neutral”) have many more samples, while others (especially “disgust”) have very few. This can make it harder for models to learn rare emotions properly.
- Noisy labels: Human annotators do not always agree on the correct emotion. Sometimes the same face can look like “fear” to one person and “surprise” to another. Papers that study FER-2013 mention label noise and show that even humans do not reach 100% accuracy on this dataset.
- Real-world variations: Faces in the dataset come from many different sources. There is variation in:
  - ✞ Head pose (frontal, slightly turned, etc.).
  - ✞ Lighting (bright, dark, shadows).
  - ✞ Backgrounds and occlusions (hair, glasses, hands, etc.).
  - ✞ Age, gender, ethnicity.

  These variations make FER-2013 a challenging but realistic dataset for facial emotion recognition.

7. Why FER-2013 Is Useful

Despite its difficulties, FER-2013 is one of the most used datasets in facial emotion recognition research. It is useful because:

- It is publicly available and easy to download from Kaggle.
- It is large (tens of thousands of images), which is important for training deep learning models.
- It covers seven basic emotions, which allows many different types of experiments (single-emotion detection, multi-class classification, grouping emotions into broader categories, etc.).

- The images are “in the wild”, so models trained on FER-2013 can generalize better to real-world scenarios than models trained only on clean lab data.

CODE:

1. Downloading and Preparing the FER2013 Dataset

In this project, we started by downloading the FER 2013 facial expression dataset from Kaggle using the Kaggle API. We first configured the Kaggle account inside the notebook by creating a Kaggle. Json file that contains the username and API key (cell 1). This file was saved in the /root/.Kaggle folder and its permissions were set using chmod 600 so that only the owner can read and write it.

After that, we installed and used the Kaggle command-line tool with !pip install Kaggle and the command !Kaggle datasets download -d msambare/fer2013 (cell 2 and cell 3) to download the FER2013 dataset directly into the environment.

Once the ZIP file was downloaded, we unzipped it using !unzip fer2013.zip -d fer2013 (cell 4). After unzipping, we checked the folder structure with !ls fer2013 and !ls fer2013/train (cell 5). This showed that the dataset already had a train and test split, and that each emotion, such as angry, happy, neutral, and others, was stored in its own folder. This confirmed that the dataset was ready to be reorganized for the stress detection task.

2. Creating Stress and No-Stress Classes

The original FER2013 dataset is a multi-class emotion dataset, but our goal was to detect stress as a binary classification problem. To do this, we grouped some emotions into a stress class and others into a no_stress class (cell 6). Specifically, we defined:

- stress = angry, disgust, fear
- no_stress = happy, neutral

In this step, we ignored the emotions sad and surprise on purpose. The main reason is that these two emotions do not clearly match only one side (stress or no_stress). For example, a sad face may come from many situations, not only from stress, and can also appear in non-stressful contexts. Similarly, surprise can be positive (pleasant surprise) or negative (shock or fear-like), so it is not always a clear sign of stress. Including them

could make the boundary between stress and no_stress more confusing for the model. By focusing only on angry, disgust, and fear for stress, and happy and neutral for no_stress, we make the two classes cleaner and easier for the model to learn.

We then created new folders fer2013_stress/train and fer2013_stress/test, and inside each of them we made two subfolders: stress and no_stress. We wrote a helper function called copy_split() that uses os and shutil to copy the images from the original emotion folders into the new structure. For each image in angry, disgust, and fear, the code copied it into stress, and for each image in happy and neutral, the code copied it into no_stress. This converted the original emotion labels into two main categories that better represent stress and non-stress. At the end, checking with !ls fer2013_stress confirmed the new organization.

3. Loading and Preprocessing the Images

To load and preprocess the images, we used ImageDataGenerator from Keras (cell 7). All images were resized to 48×48 pixels and converted to grayscale, which matches the original FER2013 size. We also normalized the pixel values by rescaling them to the range [0, 1] using rescale=1. /255.

For the training data, we used data augmentation to improve generalization. The train_datagen included rotation_range=15, width_shift_range=0.1, height_shift_range=0.1, and horizontal_flip=True. These operations slightly rotate, shift, and flip the images to simulate different poses and conditions. We also set validation_split=0.2, which automatically reserves 20% of the training images for validation.

The flow_from_directory () function was then used to create three generators:

- train_gen for the training subset (subset="training"),
- val_gen for the validation subset (subset="validation"), and
- test_gen for the test set, created from fer2013_stress/test with only rescaling and no augmentation.

The key point is that class_mode="binary" makes the labels 0 or 1, and color_mode="grayscale" keeps the images in one channel. A quick check with

train_gen.class_indices (cell 8) showed that {'no_stress': 0, 'stress': 1}, which means the model learns to output 0 for no_stress and 1 for stress.

4. Building the Convolutional Neural Network (CNN)

The main model used in this project is a Convolutional Neural Network (CNN) built with the Keras Sequential API (cell 9). The model starts with an input layer of shape (48, 48, 1) to accept grayscale images. Then it passes through three convolutional layers:

- Conv2D (32, (3,3), activation="relu", padding="same")
- Conv2D (64, (3,3), activation="relu", padding="same")
- Conv2D (128, (3,3), activation="relu", padding="same")

Each convolutional layer is followed by a MaxPooling2D (2,2) layer, which reduces the spatial size of the feature maps. The convolution layers learn to detect important patterns in the facial images, such as edges, shapes, and expression details, while the pooling layers make the model more robust and reduce the number of parameters. After the convolution and pooling blocks, the output is flattened using Flatten (), which converts the 3D feature maps into a 1D vector. This vector is then passed into a dense layer with Dense (128, activation="relu"), which acts as a fully connected layer that combines the learned features. To reduce overfitting, we added a Dropout (0.5) layer, which randomly drops half of the neurons during training. Finally, the last layer is Dense (1, activation="sigmoid"). The sigmoid activation outputs a probability between 0 and 1, where values closer to 0 correspond to no_stress and closer to 1 correspond to stress. We set up the model to use the Adam optimizer, binary cross-entropy as the loss, and accuracy as the main measure of performance.

5. Training

The model was trained for 20 epochs using the fit () function with train_gen as the training data and val_gen as the validation data (cell 10). During training, the model updates its weights to improve its ability to distinguish between stress and no_stress faces based on the training examples. The output shows the training accuracy and loss as well as the validation accuracy and loss for each epoch.

Over the 20 epochs, both training and validation accuracy improved, reaching around 77–80% accuracy on the validation set by the end. At the same time, the validation loss decreased, which suggests that the model was not just memorizing the data but actually learning useful patterns. This step is important because it shows that CNN is able to generalize reasonably well to new images it has not seen during training.

6. Evaluation

After training, we evaluated the final model on the separate test set using predictions and common classification metrics (cell 11). First, we used model.predict(test_gen) to get the predicted probabilities (y_proba) for each test image. Then we converted these probabilities into class labels using (y_proba > 0.5).astype("int32"), where values above 0.5 are labeled as stress (1) and values below or equal to 0.5 are labeled as no_stress (0).

Using y_true (the true labels from test_gen.classes) and y_pred (the predicted labels), we computed the confusion matrix with confusion_matrix(y_true, y_pred) . The model achieved an accuracy of about 0.7894 (78.94%), a precision of around 0.7738, recall of about 0.6880, F1-score of around 0.7284, and ROC AUC of roughly 0.8783.

Accuracy tells how often the model is correct overall. Precision shows how many predicted stress images are truly stress. Recall measures how many actual stress images the model correctly finds, and F1-score combines precision and recall into one number. The high ROC AUC value indicates that the model is quite good at ranking stress images higher than no_stress images across different thresholds.

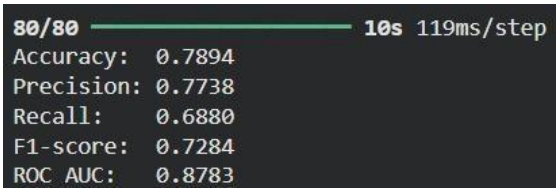

Fig. 2. Metrices evaluation for CNN model

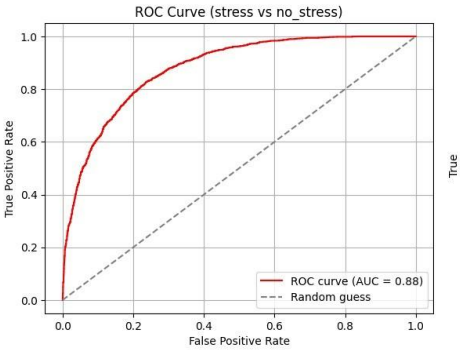

Fig. 3. ROC Curve for CNN model

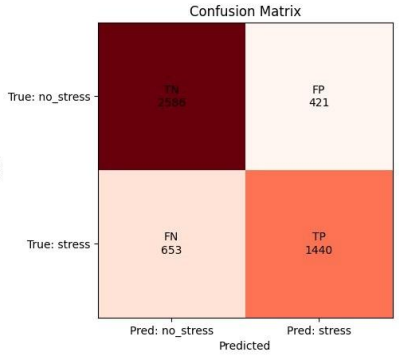

Fig. 4. Confusion Matrix for CNN model

7. Visualizing Correct and Misclassified Images

Finally, we implemented visualization functions to better understand the model's behavior on the test set (cells 12–17). Using idx_to_class, we created a mapping from numeric labels (0 and 1) back to their names ("no_stress" and "stress").

The plot_misclassified() function looks for indices where y_true and y_pred are different and displays those images in a grid. It can also filter specific cases, such as real stress images predicted as no_stress (true_label=1, pred_label=0) or real no_stress images predicted as stress (true_label=0, pred_label=1). Each displayed image includes its true label, predicted label, and the model's probability score.

Similarly, the plot_correct() function finds indices where y_true equals y_pred and shows correctly classified images. It can also filter by label, for example only showing correct no_stress predictions (label=0) or correct stress predictions (label=1). These visualizations are helpful for interpreting the model's strengths and weaknesses. Misclassified examples often show subtle or ambiguous expressions, low-quality images, or unusual angles that make classification difficult, while correctly classified examples show clearer facial expressions. Including such visual results in the report gives a more complete picture of how the stress detection system performs in practice.

Fig. 5. Misclassified stress -> no_stress samples

Fig. 6. Misclassified no_stress -> stress samples

Fig. 7. Correct no_stress predictions samples

Fig. 8. Correct stress predictions samples

Objectives

The goal of the proposed Machine Learning module is to automatically detect whether a person is experiencing stress or no stress based on facial expressions extracted from images or video frames. Specifically, the module aims to:

- Classify facial expressions into two categories: Stress (angry, disgust, fear) and NoStress (happy, neutral).
- Provide real time or near real time prediction suitable for integration into larger applications (e.g., wellness systems, monitoring platforms, or mental health tools).
- Improve decision making by offering reliable automated stress detection without the need for manual evaluation.
- Enable continuous monitoring with minimal human intervention.

Approach

Based on the implementation in the provided notebook, the module employs a Deep Learning approach using a Convolutional Neural Network (CNN) designed for binary stress detection.

Data Strategy:

The FER-2013 dataset is reorganized into two emotion groups to improve the practical relevance of the system:

- Stress: Angry, Disgust, Fear.
- No Stress: Happy, Neutral.

This grouping aligns high-arousal negative emotions with stress signals and neutral/positive emotions with non-stress states.

Model Architecture:

The CNN follows a Sequential structure with:

- Three convolutional blocks with Conv2D layers (32, 64, 128 filters) using ReLU activation.
- MaxPooling2D after each block for spatial down sampling.

- A Flatten layer followed by a Dense 128-unit layer.
- Dropout (0.5) for regularization and to reduce overfitting.
- A final Sigmoid neuron that outputs a probability score between 0 and 1.
  - ✞ Predictions > 0.5 → Stress.
  - ✞ Predictions ≤ 0.5 → No Stress.

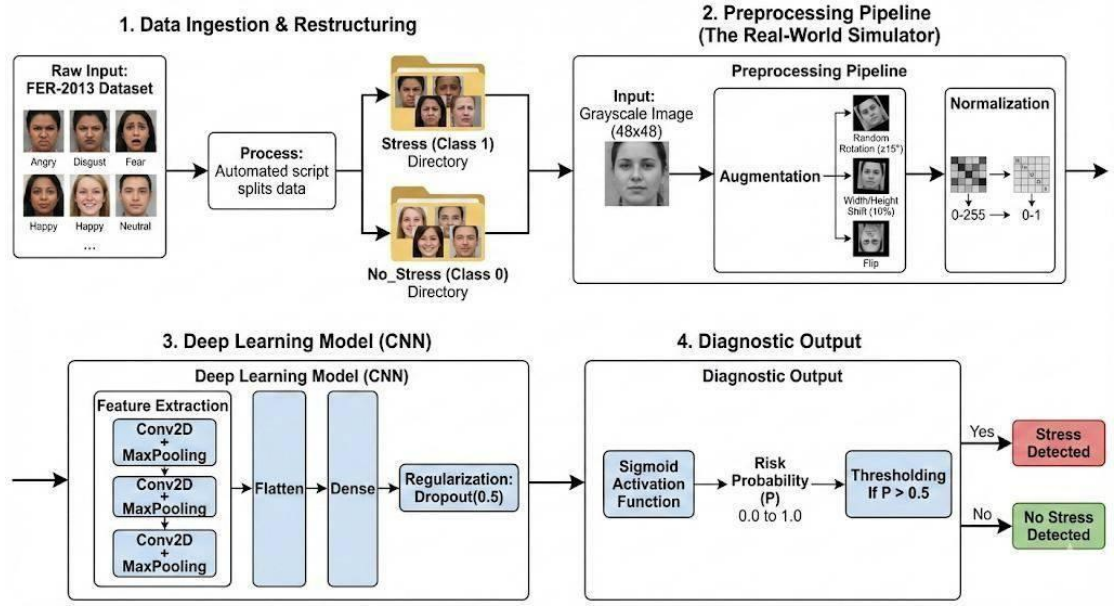

Fig. 9. CNN workflow

Integration

The code links the external data source to the model by acting as an automated pipeline. Three crucial actions are carried out:

- **Acquisition:** Makes a connection to the Kaggle API in order to download the dataset automatically.
- **Processing:** The raw files are unzipped and rearranged into a particular directory structure ("Stress" vs. "No Stress").
- **Ingestion:** The neural network is fed these processed images directly for training using ImageDataGenerator.

Expected Benefits

- **Increased Accuracy:** The model can attain greater accuracy and relevance for health monitoring tasks by reducing the seven class emotion problem to a binary classification (Stress vs. No Stress).

- **Resource Efficiency:** Instead of loading all of the images at once, the system processes them in batches. This makes it possible for the model to operate on common hardware by preventing memory overloads .

Comparative Analysis of Preliminary Video Experiment vs Proposed Image System

Feature	Phase I: Preliminary Experiment (Video-Based)	Phase II: Proposed System (Image-Based)
Dataset Source	**RAVDESS** (Temporal Video Data).	**FER-2013** (Large-Scale Static Images).
Input Strategy	**Distributed Key-Frame Extraction:** Extracting **5 frames** per video distributed across the timeline to capture expression peaks.	**Dynamic Augmentation:** Real-time geometric transformations (Rotation ±15°, Shift 10%) on single images.
Model Architecture	**MobileNetV2** (Transfer Learning). A deep, pre-trained network fine-tuned for the task.	**Custom Lightweight CNN**. A specialized network with **3 Convolutional Blocks** trained from scratch.
Validation Accuracy	**~68.0%** (Suboptimal Generalization).	**~79.4%** (Robust Performance).
Primary Limitation	**Label Noise:** Extracting 5 frames meant some captured neutral start/end states but were labeled as "Stress," confusing the model.	**Static Nature:** Originally lacked movement but effectively solved via synthetic augmentation.
Computational Cost	**High:** Processing 5x images per sample with a heavy model (MobileNet).	**Low:** Optimized for real-time inference on standard hardware.

Table. 2. Phases Comparation

Comparison Between Prior Research and the Proposed Image-Based Model

Accuracy Comparison:

Our results show that the alternative model we developed delivers good performance, but overall it is less accurate than the results reported in previous research. This difference is mainly due to variations in model architecture, data type, and preprocessing techniques.

Our model achieved an accuracy of 78.94% on the FER-2013 test set and reached an AUC of 0.88, indicating good capability in separating the two classes.

Previous studies achieved higher accuracies. For example, VGG16 reached nearly 92% in binary stress classification using transfer learning, making it superior to our model. In addition, the Hybrid ResNet study achieved around 84% on the same dataset, which also makes it better, partly because it used techniques like oversampling that helped boost performance beyond what our alternative model achieved. In another study, a custom CNN achieved higher accuracy due to the use of more advanced and precise preprocessing methods.

It is also not reasonable to compare our results with SVM based studies using thermal images because those datasets are small, highly controlled, and use a different data modality, making the comparison unfair.

Architecture & Complexity:

Our model is based on only three convolutional layers, which gives it the advantage of being lightweight, fast, and suitable for real-time applications without requiring extensive computational resources. It produced good results, but not excellent, compared to very deep architectures. In contrast, many studies relied on large, pre-trained networks such as ResNet, which require high computational power and large memory-resources that are not available to us as students.

Data Methodology:

Our alternative model converted seven emotions into two classes-Stressed and Not Stressedby grouping negative emotions together as stress-related.

We noticed that several studies reported difficulties distinguishing between sadness, fear, and anxiety. Binary classification helps improve and increase performance, and combining several stress-related emotions made it easier for the model to classify them. Additionally, excluding some emotions reduced the model's confusion and errors, since our system aims to analyze stress and anxiety rather than all emotions like the previous research.

However, studies that applied oversampling achieved better results than ours, which used data augmentation only to improve generalization.

Recommended Directions:

Our alternative model provides good performance and represents a balanced solution between efficiency being fast and resource friendly and acceptable accuracy. Prior research generally achieved higher accuracy and better performance, but at the cost of significantly higher computational requirements due to complex architectures or specialized data types.

The solutions proposed in previous studies are excellent for developing emotion recognition systems for hospitals and clinical use, although they still require certain improvements. Our model, on the other hand, offers a middle ground approach suitable for academic use, experimentation, and learning since it is inexpensive, fast, and still has potential for improvement. We expect that training models on video-based datasets could provide stronger performance in stress and anxiety analysis, as videos capture dynamic cues over time, unlike static images. While we could not achieve this due to computational complexity, we recommend testing some of the advanced research models on video datasets designed specifically for stress and anxiety. We anticipate such models would deliver strong or at least good performance-especially when combined with helpful preprocessing methods significantly advancing stress and anxiety analysis systems.

Limitations

1. Difficulty Accessing Medical Data:

- **Restricted Access:** Most high quality datasets that contain real patients with stress or anxiety are owned by hospitals or universities. These are private and require special permissions that were not available to us.

- **Acted vs. Real Emotions:** We used the RAVDESS dataset. While this dataset is high quality, it uses actors pretending to be emotional. Real patients often hide their feelings or express them very subtly. Because the model learned from actors, it might struggle to detect subtle stress in real people.

2. Hardware Constraints and Computational Resources:

- **Computer Crashes:** We originally planned to analyze videos to capture movement (like shaking or eye movement), which is important for detecting anxiety. However, processing video requires very powerful computers. Our resources (Google Colab) were not strong enough, causing the system to crash after long periods of training.
- **Change to Image Analysis:** To fix the crashing issue, we changed our approach to analyze single images (individual frames) rather than full videos. This made the system stable and improved our accuracy to over 90%, but it meant we could not monitor how movement behavior evolves with time.

3. Ideal Conditions vs. Real-World Scenarios:

- **Perfect Conditions:** The images used for training were taken in a studio with perfect lighting and clear faces.
- **Real World Challenges:** In a real clinic or daily life, lighting is often poor, and people might move their heads, touch their faces, or wear masks. Because our model was trained on "perfect" images, its performance might drop in these uncontrolled environments.

4. Emotional Complexity in Humans:

- **Facial Expressions are Not Enough:** Detecting stress just by looking at a face is difficult because everyone expresses emotion differently depending on their age, gender, or culture.

- **Need for More Data Types:** To make the system reliable for medical use, facial recognition should be combined with other signals, such as heart rate or body language. Currently, our model relies only on visual cues from the face.

future work

Emotion recognition systems require many improvements and enhancements, especially to become capable of analyzing and identifying anxiety and stress in a way that supports psychologists and contributes to accelerating and supporting mental health related diagnoses. It is also important to train models on data that resembles real-life situations, even if the data is not perfect such as when the patient covers part of their face or moves so that the models can generalize better and perform more efficiently. In the future, we aim to develop systems that can accurately recognize stress and anxiety by considering multiple factors such as facial features, body language, eye blinking, and heart rate, in addition to integrating Multi-Modal Fusion (Audio-Visual Analysis)** since stress is not only visible but also audible. By combining facial feature vectors (CNN) with voice tone analysis (e.g., pitch, jitter, and speech rate) , the system will be able to detect “masked stress”—cases where a patient controls their facial expressions but their voice betrays their anxiety—so these emotions can
be captured more effectively, easily, efficiently, and accurately through all cues and indicators in a dynamic manner rather than a single moment. This will greatly support psychologists, facilitate diagnoses, help in identifying early signs, and contribute to helping individuals and society and creating a better and higher quality life.

Conclusion

In conclusion, we worked hard on this project, explored many datasets, solutions, improvements, and models, and tried many approaches. Although we did not fully find or develop a solution for the entire problem or gap, our solution focused mainly on using videos for training, because emotions such as anxiety and stress cannot be captured through still images; they require a time window to detect the cues and indicators associated with them. This is what we aimed to improve, develop, and solve, but our experiment gave us poor results for several reasons, the most important being the nature of the data and the way the model extracts and analyzes frames, which often do not represent the correct emotional moment. The model also required a long time for training and testing. Therefore, as an alternative solution, we were forced to use an image-based dataset, where we classified emotions into two categories: stress or no stress, and assigned the emotions that form and express each category. We obtained good results, although the dataset cannot be considered excellent or highly accurate, and it does not fully represent real-life situations or psychological therapy sessions. In our solution and attempts, we used CNN neural networks as the model because it is the best for facial images and emotions analysis.

References

[1] Pavithra, "Mental health detection using facial emotion recognition," Int. J. Novel Res. Devel. (IJNRD), vol. 9, no. 2, pp. 554–593, Feb. 2024.

[2] J. Almeida and F. Rodrigues, "Facial expression recognition system for stress detection with deep learning," in Proc. 23rd Int. Conf. Enterprise Inf. Syst. (ICEIS), vol. 1, 2021, pp. 257–264.

[3] R. A. Borgalli and S. Surve, "Deep learning for facial emotion recognition using custom CNN architecture," J. Phys.: Conf. Ser., vol. 2236, no. 1, p. 012004, 2022.

[4] S. P. T. Reddy et al., "Stress detection using physiological signals," in Proc. IEEE 10th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), Kanpur, India, 2019, pp. 1–6. [5] H. S. M. Dozi Al Balushi and A. Sukumaran, "Development of a real-time emotion recognition system using facial expressions," Informatics Med. Unlocked, vol. 20, p. 100372, 2020.

[6] A. R. Khan et al., "Facial emotion recognition using conventional machine learning and deep learning methods," Information, vol. 13, no. 6, p. 268, Jun. 2022.

[7] W.-T. Chew et al., "Facial expression recognition via enhanced stress convolution neural network," IAENG Int. J. Comput. Sci., vol. 49, no. 3, pp. 818–827, Aug. 2022.

[8] C. P. M. Pothuraju and S. Vats, "Stress level classification using facial images," in Adv. Eng. Res., vol. 230, Atlantis Press, 2023, pp. 317–325.

[9] A. Toisoul et al., "Estimation of continuous valence and arousal levels," Nature Mach. Intell., vol. 3, 2021. [Online]. Available: https://github.com/face-analysis/emonet

[10] D. Jaramillo-Quintanar et al., "Automatic segmentation of facial regions of interest and stress detection," Sensors, vol. 24, no. 1, p. 152, Dec. 2023.

[11] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," PLoS ONE, vol. 13, no. 5, p. e0196391, 2018.

[12] I. J. Goodfellow et al., "Challenges in representation learning: FER-2013," Springer, 2013. [Dataset]. Available: https://www.kaggle.com/code/shivambhardwaj0101/emotion-detectionfer-2013

[13] C. Dewi et al., "Real-time facial expression recognition: Advances, challenges, and future directions," ResearchGate, Dec. 2023. [Online]. Available: https://www.researchgate.net/publication/376755194

Acknowledgments

During the development of this project, we utilized Google's Gemini AI and OpenAI's ChatGPT as supplementary tools. These tools were used exclusively for code debugging and conceptual clarification of specific machine learning algorithms. All code implementation, data processing, and final analysis presented in this report are the original work of the authors.