**DEVRELAX: STRESS MONITORING AND RELIEVING APPLICATION FOR IT PROFESSIONALS**

Janudi Disara Ranasinghe

IT20037888

BSc (Hons) in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2023

**DEVRELAX: STRESS MONITORING AND RELIEVING APPLICATION FOR IT PROFESSIONALS**

Janudi Disara Ranasinghe

IT20037888

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Hons) in Information Technology Specializing in Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology

Sri Lanka

September 2023

# DECLARATION

I declare that this is my own work and this dissertation1 does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

A close up of a name

Description automatically generated

Signature: Date: 08/09/2023

The above candidate has carried out research for the bachelor’s degree Dissertation under my supervision.



Signature of the supervisor: Date: 08/09/2023

(Mr. Samadhi Rathnayake)

# ABSTRACT

This research delves into the domain of stress detection through facial dynamics, with a particular focus on IT professionals. Stress is a pervasive issue in modern workplaces, affecting productivity and well-being. While facial expression analysis has shown promise as a means to identify stress, this study seeks to enhance its accuracy by considering the usage of attention mechanisms. This innovative application harnesses the power of machine learning to discern and categorize stress levels, drawing insights from a multitude of sources. These include data gleaned from a heart rate sensor integrated with the mouse and keyboard, as well as nuanced facial expressions. The proposed solution unfolds in two integral components: stress monitoring and stress relieving. The former meticulously collects data from the sensors, scrutinizing usage patterns to predict stress levels accurately. The latter facet extends personalized recommendations and activities, tailored to alleviate stress levels commensurate with the user's recognized stress level. Facial expressions serve as valuable indicators of mental states, including stress, but the manifestation of stress can vary significantly among individuals. This study addresses this variability by developing a facial expression analysis algorithm that takes attention mechanisms into account. By wielding the prowess of machine learning techniques and offering individualized recommendations, this application is poised to have a profound impact on the mental well-being and productivity of IT professionals. By bridging the gap in stress detection through facial dynamics and considering individual characteristics, this study offers a promising avenue for improving the well-being and productivity of IT professionals in demanding work environments.

Keywords: Deep Learning algorithms, Deep Neutral Net Binary Classifier, Visual Transformers (ViT), Convolutional Neural Networks (CNN)

# ACKNOWLEDGMENT

I am deeply grateful for the unwavering support and guidance provided by my dissertation supervisor, Mr. Samadhi Rathnayake, and co-supervisor, Ms. Devanshi Ganegoda, from the Faculty of Computing.

I extend my heartfelt thanks to the dedicated staff of the Faculty of Computing for their consistent assistance and motivation throughout my academic journey. Their commitment to both teaching and research has been a true source of inspiration. Furthermore, I would like to thank our lecturer in charge, Dr. Jayantha Amararachchi and the CDAP team for their assistance in ensuring the project's success.

I also wish to express my appreciation to my fellow team members for their invaluable support and collaborative efforts. Our insightful discussions and idea exchanges have significantly enriched my comprehension of the subject matter.

Special gratitude is owed to Mrs. Arosha Dassanayake, our external supervisor specializing in psychology, for her generous support in the domain of stress-related research. Her contributions have been instrumental in the successful realization of this project.

Lastly, I extend my thanks to the CEO of Xinotech for their invaluable support in the commercialization of our application. Their backing has been instrumental in translating our research into a practical and impactful solution.

Finally, I would like to express my appreciation to all those who participated in this study and contributed their time and insights. Their willingness to share their experiences and perspectives has been invaluable in helping me to gain a deeper understanding of the subject matter.

# TABLE OF CONTENTS

[DECLARATION 3](#_Toc145284203)

[ABSTRACT 4](#_Toc145284204)

[ACKNOWLEDGMENT 5](#_Toc145284205)

[TABLE OF CONTENTS 6](#_Toc145284206)

[LIST OF FIGURES 7](#_Toc145284207)

[LIST OF TABLES 9](#_Toc145284208)

[LIST OF ABBREVIATIONS 10](#_Toc145284209)

[1. INTRODUCTION 11](#_Toc145284210)

[1.1 Background Literature 11](#_Toc145284211)

[1.2 Research Gap 15](#_Toc145284212)

[1.3 Research Problem 18](#_Toc145284213)

[1.4 Research Objectives 20](#_Toc145284214)

[1.4.1 Main Objective 20](#_Toc145284215)

[1.4.2 Specific Objectives 21](#_Toc145284216)

[2. METHODOLOGY 23](#_Toc145284217)

[2.1 Methodology 23](#_Toc145284218)

[2.1.1 Requirement Gathering 23](#_Toc145284219)

[2.1.2 Dataset Description 24](#_Toc145284220)

[2.1.3 System Architecture 24](#_Toc145284221)

[2.1.4 Component Architecture 25](#_Toc145284222)

[2.1.5 Tools and Libraries 26](#_Toc145284223)

[2.1.6 Model Architecture 26](#_Toc145284224)

[2.2 Commercialization Aspects Of The Product 44](#_Toc145284225)

[2.2.1 Target Market 44](#_Toc145284226)

[2.2.2 Revenue Streams 44](#_Toc145284227)

[2.2.3 Marketing Approach 44](#_Toc145284228)

[2.3 Testing and Implementation 46](#_Toc145284229)

[2.3.1 Functional Requirements 46](#_Toc145284230)

[2.3.2 Non-Functional Requirements 46](#_Toc145284231)

[2.3.3 Backend Implementation 47](#_Toc145284232)

[2.3.4 Backend Testing 54](#_Toc145284233)

[3. RESULTS AND DISCUSSION 56](#_Toc145284234)

[3.1 Results 56](#_Toc145284235)

[3.1.1 Convolutional Neural Network (CNN) Model Results 56](#_Toc145284236)

[3.1.2 Visual Transformer (ViT) Model Results 57](#_Toc145284237)

[3.2 Research Findings 59](#_Toc145284238)

[3.3 Discussion 60](#_Toc145284239)

[3.3.1 Model Performance and Comparative Analysis 60](#_Toc145284240)

[3.3.2 Practical Implications and Applications 60](#_Toc145284241)

[3.3.3 Limitations and Future Directions 61](#_Toc145284242)

[3.3.4 Ethical Considerations 62](#_Toc145284243)

[3.4 Summary of Each Student’s Contribution 62](#_Toc145284244)

[4. CONCLUSION 64](#_Toc145284245)

[5. REFERENCES 67](#_Toc145284246)

[6. APPENDIX 69](#_Toc145284247)

# LIST OF FIGURES

[Figure 1: Causes of workplace stress 11](#_Toc145282108)

[Figure 2: Overview of the system diagram 24](#_Toc145282109)

[Figure 3: Overview of the component diagram 25](#_Toc145282110)

[Figure 4: Import libraries for pre-processing 26](#_Toc145282111)

[Figure 5: Object creation class 26](#_Toc145282112)

[Figure 6: Split data into training and testing 27](#_Toc145282113)

[Figure 7: Convert string of pixel values to images 27](#_Toc145282114)

[Figure 8: Save images to directories 27](#_Toc145282115)

[Figure 9: Libraries imported for the data loading task 28](#_Toc145282116)

[Figure 10: Plain\_Dataset class implementation 28](#_Toc145282117)

[Figure 11: Helper function implementation 29](#_Toc145282118)

[Figure 12: Spatial attention mechanism implementation 29](#_Toc145282119)

[Figure 13: Hierarchical attention mechanism implementation 30](#_Toc145282120)

[Figure 14: Deep Emotion class implementation I 31](#_Toc145282121)

[Figure 15: Deep Emotion class implementation II 32](#_Toc145282122)

[Figure 16: Import the required libraries 33](#_Toc145282123)

[Figure 17: Train function 33](#_Toc145282124)

[Figure 18: Training the model through using Deep\_Emotion class 34](#_Toc145282125)

[Figure 19: Importing the necessary libraries 38](#_Toc145282126)

[Figure 20: Extracting the pixels and labels from the dataset 38](#_Toc145282127)

[Figure 21: Converting pixels and stacking images 39](#_Toc145282128)

[Figure 22: Splitting the dataset and augmentation applied 39](#_Toc145282129)

[Figure 23: LBP and PCA applied 39](#_Toc145282130)

[Figure 24: Building the ViT model 40](#_Toc145282131)

[Figure 25: Imports and blueprint definition 46](#_Toc145282132)

[Figure 26: Deep emotion model initialization 47](#_Toc145282133)

[Figure 27: Function to prompt camera and save image 47](#_Toc145282134)

[Figure 28: API implementation I 48](#_Toc145282135)

[Figure 29: API implementation II 48](#_Toc145282136)

[Figure 30: API implementation III 49](#_Toc145282137)

[Figure 31: Redux slice 49](#_Toc145282138)

[Figure 32: Face Recognition Reducer 49](#_Toc145282139)

[Figure 33: JavaScript code implementations for camera 50](#_Toc145282140)

[Figure 34: Backend route call 50](#_Toc145282141)

[Figure 35: EmotionPred API 51](#_Toc145282142)

[Figure 36: Postman for testing the backend APIs 52](#_Toc145282143)

[Figure 37: Dashboard 52](#_Toc145282144)

[Figure 38: Prompted camera interface 53](#_Toc145282145)

# LIST OF TABLES

[Table 1: Previous Related Research Work 17](#_Toc145282429)

[Table 2: Research Gap 18](#_Toc145282430)

[Table 3: Confusion matrix (CNN) 55](#_Toc145282431)

[Table 4: Confusion Matrix (ViT) 56](#_Toc145282432)

[Table 5: Student contributions 62](#_Toc145282433)

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Long Form** |
| CNN | Convolutional Neural Network |
| ViT | Vision Transformer |

# INTRODUCTION

## Background Literature

As the landscape of work continues to evolve, with remote work becoming increasingly prevalent, a new set of challenges has emerged, particularly within the IT industry [1]. This shift towards remote work, while offering flexibility and autonomy, has also led to a concerning trend: the isolation of employees, potentially resulting in heightened levels of stress and related health issues.

This issue is poignantly exemplified by the tragic case of Matsuri Takahashi, a young employee who, in the pursuit of her career, tragically lost her life due to extreme stress and overwork. Investigation following her untimely death revealed a direct link between her work conditions and the extreme stress she faced [2]. This tragic incident serves as a somber reminder of the critical importance of safeguarding the mental well-being of professionals in high-pressure work environments.

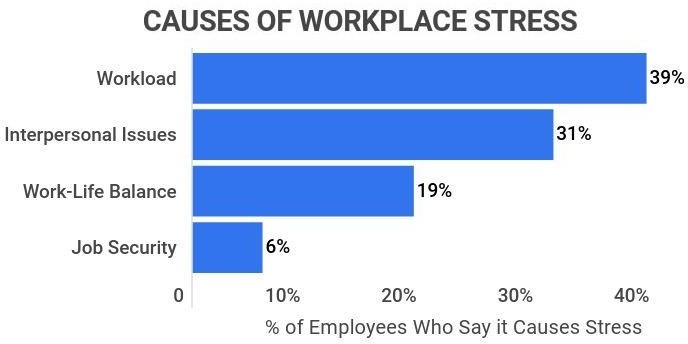


Figure 1: Causes of workplace stress

Stress, as a complex and multifaceted phenomenon, can lead to an array of adverse consequences, ranging from mental health issues and strained interpersonal relationships to severe cases of depression and, tragically, even suicide. While complete eradication of stress may be an unattainable goal, the implementation of preventive measures is paramount in alleviating its impact.

Presently, stress detection predominantly relies on symptom-based assessments that are largely dependent on self-reporting. This approach, however, poses significant challenges in accurately gauging one's stress levels. It often necessitates the intervention of medical and physiological experts, making it an impractical and resource-intensive method for widespread stress monitoring.

However, there are tools that use machine learning algorithms to predict if a person has stress, based on physiological factors such as heart rate. The issue with these methodologies is that they are solely based on just one factor, rather than considering many aspects. In recent with the help of deep learning, many approaches to monitor stress were introduced and some of the approaches are mentioned below.

* Facial Emotion Recognition: Facial emotion recognition is a popular technique for detecting tension. Deep learning systems can be taught to identify stress-related face characteristics such as furrowed brows, tense lips, and narrowed eyes.
* Speech Analysis: Another method for detecting stress is through speech analysis. Deep learning algorithms can be taught to evaluate a person's speaking tone, intonation, and rhythm and find patterns that indicate tension.
* Wearable Devices: Smartwatches and exercise monitors, for example, can be used to identify stress. These devices can gather data such as heart rate, skin conductance, and exercise level, which can then be evaluated with deep learning algorithms to identify stress patterns

All the above methods are focused on just a single approach where either physiological signals or speech or facial recognition are considered. Therefore, just a single factor is considered and the result for stress is based solely on this considered factor. Furthermore, the above approaches are suitable for individuals and it is not applied strictly to an integrated environment, for IT professionals, as the system that is being proposed.

Recent advancements in technology, particularly in the realm of machine learning [3], have offered new avenues for stress detection through various physiological and behavioral cues. Among these, facial dynamics have emerged as a potent modality for assessing stress levels. Facial expressions, known to convey a wealth of emotional information, have long been recognized as indicators of affective states, including stress.

The intricate choreography of facial muscles in response to emotional experiences provides a rich wellspring of data that holds promise for non-intrusive, real-time stress assessment. However, a significant hurdle in crafting resilient stress detection models is the artful allocation of attention across diverse facial features, contingent on their impact on stress levels. This challenge is particularly acute when contemplating the application of hierarchical attention mechanisms, given the high variability and susceptibility to noise inherent in facial dynamics [4]. Striking the right balance between capturing salient features and dampening the influence of extraneous signals is pivotal to this pursuit.

To surmount this obstacle, our research introduces an innovative approach incorporating a dynamic hierarchical attention mechanism. This design is tailored to autonomously discern the optimal attention weights for distinct facial expressions. By bestowing varying degrees of significance to different facial zones based on their respective contributions to stress levels [5], our model amplifies the resilience and adaptability of stress detection. This pioneering methodology harbors the potential to transform stress monitoring and set the stage for more potent and dependable systems.

Continuous progress in this domain not only holds the potential to enhance applications in stress monitoring but also facilitates proactive interventions and customized stress management techniques. Furthermore, the insights gleaned from this investigation may transcend the realm of stress detection, finding relevance in areas like emotion recognition, evaluations of mental well-being, and interactions between humans and computers.

Through affording a more nuanced comprehension of emotional states, our methodology stands poised to contribute to the crafting of more resilient and trustworthy systems for real-time emotional assessment. The discoveries stemming from this research add to the expanding repository of knowledge in affective computing and chart a course for the creation of systems that harness facial dynamics for a wide spectrum of applications beyond stress detection. Given the intricate and mutable nature of facial emotions, discerning them constitutes a formidable challenge in the arena of facial recognition.

By providing a more nuanced understanding of emotional states, our approach has the potential to contribute to the development of reliable systems for real-time emotional analysis. Since facial emotions are complicated and changing, recognizing them is a difficult job in facial recognition.

Facial expressions are created by movements of the facial musculature and are typically brief, subtle, and involve numerous areas of the face. Facial expression interpretation is essential for a variety of uses, including mood identification, deception detection, and human-computer interaction.

This research represents a significant step forward in harnessing the power of facial dynamics for stress detection, offering a novel and effective approach to addressing the mental health challenges faced by individuals in demanding work environments. Through the development of innovative methodologies and the exploration of dynamic attention weighting strategies, this research stands to make a meaningful contribution to the field of affective computing, with far-reaching implications for the well-being and productivity of individuals across various domains.

## Research Gap

Conventional approaches to stress detection and counseling have inherent limitations that can be effectively addressed through the application of deep learning techniques. One of the primary drawbacks of traditional counseling lies in its subjective nature, heavily reliant on self-reported symptoms and external observations. This methodology often falls short in accuracy, as individuals may struggle to accurately gauge their own stress levels or may be hesitant to openly express their true feelings. Moreover, traditional counseling services may not be universally accessible or affordable, particularly for individuals residing in remote or resource-constrained regions.

In contrast, leveraging deep learning for stress detection offers a paradigm shift by enabling the objective and automated analysis of physiological and behavioral data. This approach holds the promise of providing assessments that are not only more accurate but also consistently reliable compared to human evaluations. Additionally, the integration of ML techniques introduces the potential for real-time monitoring of stress levels, facilitating timely interventions. This feature proves especially valuable in high-stress environments like workplaces and educational institutions. Thus, while conventional counseling methods have their merits, the application of deep learning for stress detection presents a more objective and accessible avenue with broader potential benefits.

Present models have predominantly focused on utilizing either physiological signals or facial recognition alone. This presents an opportunity to explore the potential advantages of incorporating heart rate data and keyboard dynamics alongside facial expressions and emotion recognition for a more accurate and comprehensive approach to stress detection. This gap in the literature underscores the need for further exploration to enhance the effectiveness of stress identification using facial expression analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| **Ref** | **Title** | **Dataset** | **Result** |
| [6] | Stress detection with machine learning and deep learning using  multimodal physiological data | Public dataset WESAD dataset | Achieved accuracy of 84.32% and 95.21%  using RF, DT, KNN, AdaBoost, LDA, SVM and DL |
| [7] | Stress detection through speech analysis using machine learning | Public Dataset Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset | CNN: 94.26% accuracy achieved |
| [8] | Introducing WESAD, a multimodal dataset for wearable stress and affect detection | Public dataset WESAD dataset | Accuracy of 80% (three class) and 93% (two class) was achieved using RF, DT, KNN, AdaBoost, LDA, and SVM |
| [9] | Machine learning and IoT for prediction and detection of stress | Private Dataset Collected own dataset and classified using Python | LR-66% SVM-68% |

Table 1: Previous Related Research Work

The proposed research endeavors to bridge these gaps by adopting a multifaceted approach that considers physiological signals, facial dynamics, and emotion recognition. By doing so, it aims to provide a more nuanced and comprehensive understanding of stress levels, particularly among IT professionals. The study not only focuses on stress detection but also encompasses the development of methods for stress alleviation, including the implementation of customized models utilizing reinforcement learning.

Moreover, the research aims to introduce an innovative stress detection and relief application tailored to the distinctive needs of IT workers. Furthermore, the integration of the application with common IT workflow systems will streamline its usability for IT professionals, ensuring seamless incorporation into their daily routines. This addresses a prevalent limitation in many existing stress management apps, which often function as standalone products, potentially leading to limited user engagement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Research Gap | Research A [10] | Research B  [11] | Research C  [12] | Proposed Solution |
| Not solely focused on facial dynamics | No | No | No | Yes |
| Use of attention mechanisms for selective focus | No | No | No | Yes |
| Using a frame to detect face | Yes | Yes | Yes | Yes |
| Usage of facial landmarks | Yes | Yes | No | Yes |
| Categorizes to different emotion levels | Yes | Yes | Yes | Yes |
| Gives a weight for the stress level for each emotion | No | Yes | Yes | Yes |
| Integrated into a workflow environment | No | No | No | Yes |
| Addresses the issue of overfitting | No | No | No | Yes |

Table 2: Research Gap

In summary, this research aims to fill critical research gaps in the field of stress detection and relief for IT professionals by adopting a comprehensive and personalized approach. Through the integration of physiological signals, facial dynamics, and emotion recognition, coupled with the development of a tailored application, this study holds the potential to significantly impact the well-being and productivity of IT professionals in high-pressure work environments.

## Research Problem

In today's relentless, high-pressure environment, stress has evolved into a pervasive concern, significantly impacting individuals' well-being and overall quality of life. The development of robust stress detection systems stands as a critical endeavor, enabling proactive interventions and personalized stress management strategies. This research endeavors to leverage facial dynamics as a modality for stress detection, employing state-of-the-art techniques like Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). However, while considerable strides have been made in mitigating overfitting and noise in facial dynamics analysis, a more refined exploration of attention mechanisms is warranted.

One of the primary challenges in developing robust stress detection models lies in optimizing dynamic hierarchical attention mechanisms. These mechanisms play a pivotal role in directing the model's focus towards essential facial features. However, the intricate and variable nature of facial dynamics, often susceptible to noise, necessitates mechanisms with adept selectivity. This research problem calls for the fine-tuning of these attention mechanisms to navigate the complex terrain of facial dynamics effectively. The goal is to develop attention mechanisms that can adeptly capture pertinent facial features while mitigating the influence of noisy signals, thereby enhancing the robustness and generalizability of stress detection models.

Moreover, it is essential to conduct a comprehensive evaluation of the relative efficacy of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in the context of stress detection. While both architectural paradigms possess distinct advantages – CNNs' proficiency in capturing local features and ViTs’ aptitude for discerning global relationships – a nuanced understanding of their comparative strengths and limitations is crucial. This research problem strives to elucidate the optimal architectural choice for extracting salient facial features, aiming to significantly advance stress detection methodologies.

Facial dynamics serve as a rich tapestry of emotional expression, providing invaluable insights into an individual's affective state. However, the challenge lies in discerning the salient features amidst the complexity of facial movements. Current approaches often grapple with overfitting and noise, underscoring the critical need for attention mechanisms that not only capture pertinent features but also mitigate the impact of extraneous signals. The dynamic hierarchical attention mechanism proposed in this research aims to precisely address this challenge, offering a refined approach to stress detection through facial dynamics.

In the realm of facial dynamics analysis, the application of attention mechanisms introduces a notable conundrum – the risk of overfitting. Overfitting occurs when a model becomes excessively attuned to the idiosyncrasies of the training data, to the extent that it struggles to generalize effectively to unseen data. When attention mechanisms are applied without due consideration, there is a tendency for the model to assign undue significance to noise and irrelevant features within the facial dynamics.

The complexity of facial dynamics, often characterized by subtle nuances and interplays of muscle movements, poses a significant challenge in stress detection. Attention mechanisms, while instrumental in highlighting relevant facial features, must tread carefully to avoid amplifying the impact of extraneous signals. Without refined attention mechanisms, the model risks allocating disproportionate weight to inconsequential variations in facial expressions, thereby compromising the accuracy of stress assessments.

A pivotal facet of this research problem centers on the optimization of attention mechanisms. These mechanisms hold the key to guiding the model's focus towards critical facial features that encapsulate stress-related cues. However, the standard application of attention mechanisms may inadvertently exacerbate the challenges posed by the inherent variability and noise within facial dynamics. The goal, therefore, is not merely to implement attention mechanisms, but to fine-tune them with precision. This fine-tuning process involves honing the model's ability to discriminate between relevant and irrelevant facial features, ultimately enhancing the robustness and generalizability of stress detection models.

To further complicate matters, the hierarchical nature of facial dynamics necessitates a multi-level attention strategy. Different regions of the face contribute distinctively to the overall expression of stress. Conventional attention mechanisms often fall short in accommodating this granularity, potentially leading to suboptimal stress assessments. This research problem calls for the development of attention mechanisms that are not only dynamic but also hierarchical in nature. By enabling the model to discern the relative importance of different facial regions, the attention mechanisms can effectively capture the nuanced interplay of features critical to stress detection.

## Research Objectives

### 1.4.1 Main Objective

Stress, an omnipresent companion in the lives of IT professionals, often silently accumulates during long hours of work, leaving a trail of potential health issues in its wake. The overarching objective of this research is to offer a holistic solution to tackle this pervasive issue. By developing a comprehensive stress detection and relief system tailored for IT workers, this research seeks to address not only the identification of stress levels but also the provision of effective stress reduction strategies. The core aim revolves around amalgamating multiple stress detection methodologies, including facial expression analysis, heart rate monitoring, and keyboard dynamics, to enhance precision in gauging stress levels. In tandem, the system will harness the power of machine learning techniques, notably reinforcement learning, to offer personalized recommendations for stress-alleviating activities.

### 1.4.2 Specific Objectives

**Sub-Objective 1:** Identifying Stress-Related Facial Action Units (AUs) and Developing a Hierarchical Attention Mechanism

* Specific: The research endeavors to pinpoint the facial action units (AUs) associated with stress among IT professionals. Furthermore, it aims to devise a hierarchical attention mechanism designed to accentuate crucial facial features.
* Measurable: Within the initial three months of the research, before the first progress presentation, the goal is to identify the facial AUs corresponding to stress in at least 90% of the collected facial expressions.
* Achievable: Prior studies have successfully utilized facial AUs for similar investigations, and the available facial expression dataset appears adequate for AU identification.
* Relevant: Recognizing the facial AUs linked to stress constitutes a pivotal step in creating an effective stress detection algorithm based on facial expressions.
* Time-bound: Completion of facial AU identification before the first evaluation progress presentation will ensure that the project adheres to its schedule.

**Sub-Objective 2:** Preprocessing and Preparing of Facial Expression Dataset

* Specific: Implement a comprehensive preprocessing and annotation process for the gathered facial expression dataset, optimizing it for the training of the facial expression analysis algorithm.
* Measurable: Accomplish the preparation and labeling of at least 80% of the amassed facial expression dataset prior to the first progress presentation of the research.
* Achievable: The requisite software tools and resources for data preprocessing are readily accessible, ensuring the feasibility of this objective.
* Relevant: The availability of a well-prepared and annotated facial expression dataset forms the bedrock for the development and training of the facial expression analysis algorithm.
* Time-bound: The completion of data preprocessing and labeling, scheduled prior to the first progress presentation, serves as a critical milestone to ensure project progress.

**Sub-Objective 3:** Investigation of Convolutional Neural Networks (CNNs) and Overfitting Mitigation

* Specific: Develop a sophisticated facial expression analysis algorithm leveraging advanced deep learning techniques, particularly Convolutional Neural Networks (CNNs), while considering the nuances of hierarchical attention mechanisms.
* Measurable: Complete the preparation of no less than 80% of the gathered facial expression dataset before the first progress presentation of the research, ensuring a robust foundation for algorithm development.
* Achievable: Employ a meticulous examination of existing methods to mitigate overfitting and formulate an innovative strategy tailored to this research context.
* Relevant: The strategy devised to counteract overfitting must directly address the pertinent challenge, contributing significantly to the resolution of the research problem.
* Time-bound: Concluding data preprocessing and labeling before the initial progress presentation is a strategic imperative, pivotal to maintaining the project's momentum and trajectory.

Some additional sub-objectives that can be considered for this project are also explained in brief in the following context.

**Additional Objective 1:** Explore Multimodal Integration

* Investigate the potential benefits of integrating multiple modalities, such as facial expression analysis, heart rate monitoring, and keyboard dynamics, for enhanced stress detection accuracy.
* Develop algorithms for effective fusion and interpretation of multimodal data streams.

**Additional Objective 2:** Benchmark against Existing Stress Detection Methods

* Compare the performance of the proposed system against established stress detection methods to demonstrate its superiority.
* Conduct rigorous experiments and statistical analyses to validate the system's efficacy.

**Additional Objective 3:** Develop Noise Reduction Strategies

* Investigate methods to mitigate noise and artifacts in facial dynamics data, ensuring that the attention mechanisms focus on meaningful features.
* Implement filtering and preprocessing techniques to enhance the signal-to-noise ratio in facial expression analysis.

**Additional Objective 4:** Investigate User Feedback and Preferences

* Gather feedback from IT professionals regarding their preferences for interacting with the stress detection system.
* Incorporate user-centric design principles to enhance the usability and acceptance of the facial dynamics component.

# METHODOLOGY

## Methodology

### 2.1.1 Requirement Gathering

The foundation of any research endeavor lies in meticulous requirement gathering. In this phase, the research topic and background study were rigorously examined. A comprehensive understanding of current processes and comparable systems was acquired through an extensive analysis of predefined criteria. To set the project's boundaries, the scope was meticulously specified. In the case of this application, key stakeholders, including a psychologist and an IT expert (external supervisors), were interviewed to gather additional insights. Feedback from workers in the IT sector was also solicited to ensure a well-rounded understanding of suitable system characteristics.

Key Stages:

* Collecting related research papers.
* Conducting a feasibility study.
* Conducting a background and literature assessment.
* Reading and evaluating the collected research papers.
* Gathering data from users and evaluating their perspectives on the system.
* Identifying the most suitable components and finalizing the project scope.

### 2.1.2 Dataset Description

For stress level identification, readily available datasets were meticulously curated. These datasets serve as the primary sources of inputs used to ascertain stress levels. Notably, the datasets used in this research have already been collected and are publicly accessible. Moreover, datasets pertaining to heart rate and keyboard dynamics are also publicly available.

At the core of our research lies the Fer2013 dataset, a comprehensive repository of grayscale facial images designed to encapsulate a wide array of emotional states. Each of these images is standardized to a size of 48x48 pixels through an automated registration process, ensuring uniformity and centralization. Boasting an extensive collection of approximately 30,000 images, this dataset presents a diverse spectrum of facial expressions, spanning across seven distinct emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. It's worth noting that the distribution of these emotions within the dataset exhibits some variation. Specifically, the Disgust category is represented by a more modest count of 600 images, while the remaining labels feature a more substantial presence, each accounting for nearly 5,000 samples. The Fer2013 dataset forms the indispensable foundation upon which we build, enabling the training, validation, and evaluation of stress detection models that hinge on the dynamics of facial expressions.

### 2.1.3 System Architecture

The system architecture, illustrated below, embodies the core of this research project. The primary objective is to develop a machine learning and reinforcement learning-based application for stress identification and reduction, tailored specifically for IT workers. The system aggregates user data from multiple sources, including keyboard dynamics, heart rate dynamics, and facial dynamics, as depicted in the diagram. Facial dynamics are captured using a camera that captures images of the user's face. Subsequently, a reinforcement learning method is devised to learn and offer individualized stress-relieving activities based on the user's stress level. These activities are meticulously customized to align with the user's preferences and needs, offering a highly efficient intervention for stress release.

To ensure precise stress detection and individualized stress reduction recommendations, the application's performance is assessed through thorough user testing and crossvalidation. The overall goal of the study project is to offer a more efficient method of stress management that is accurate, individualized, and catered to specific needs.

A diagram of a diagram

Description automatically generated

Figure 2: Overview of the system diagram

### 2.1.4 Component Architecture

The stress detection component of the application uses heart rate sensors, keyboard dynamics and the camera, which is prompted, to monitor the user's stress levels at 20-minute intervals. These images are preprocessed, and the pre-trained model is used to detect the facial expressions of the user, such as happy, sad, or angry. Each facial expression is assigned a probability that contributes to the overall stress level.

The output from each of the components is then aggregated using a common metric and the final stress level will be sent to the recommendation and alleviation system. This component provides a non-intrusive and accurate way to measure stress levels and can be used to inform personalized stress relief recommendations. The diagram below provides a visual representation of the stress detection via facial dynamics component of the application.

A diagram of a computer process

Description automatically generated with medium confidence

Figure 3: Overview of the component diagram

### 2.1.5 Tools and Libraries

* PyTorch: The implementation of model architecture and training pipeline leverages PyTorch, a prominent deep learning framework.
* OpenCV: Open-Source Computer Vision Library (OpenCV) is instrumental in image preprocessing, face detection, and visualization.
* Pandas: The Pandas library facilitates the handling and manipulation of CSV data.
* Matplotlib: Matplotlib is adept at generating data visualizations.

### 2.1.6 Model Architecture

We have diligently developed two distinct model architectures to address stress detection: Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). The CNN-based model is enriched with hierarchical attention mechanisms, intelligently allocating dynamic weights to facial features based on their significance. In parallel, the ViT-based model incorporates hierarchical attention mechanisms, concentrating on unique facial regions pivotal for stress detection. Both models effectively leverage the inherent strengths of their respective architectures to decode and interpret facial dynamics, consequently enabling precise stress assessment. In the initial phase, considerable attention was devoted to the pre-processing and refinement of the dataset.

**Step 1:** Initially the relevant libraries were imported for data manipulation, numerical computations, image processing, file system operations, and progress bar functionality.

A close-up of a white background

Description automatically generated

Figure 4: Import libraries for pre-processing

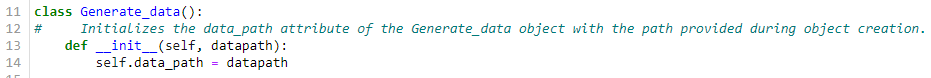
**Step 2:** A Generate\_data class, passing the appropriate data path during object creation was created. 

Figure 5: Object creation class

**Step 3:** The original file was then split into training and test files, using a function as shown below.

A computer code with many colorful text

Description automatically generated with medium confidence

Figure 6: Split data into training and testing

**Step 4:** This method converts a string of pixel values from the CSV file into a PIL image object. It can be used to preprocess the pixel values before training.

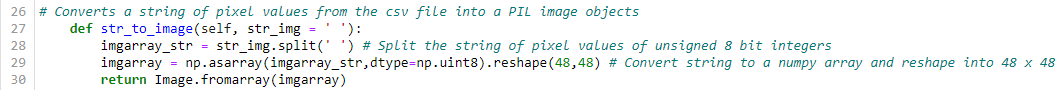


Figure 7: Convert string of pixel values to images

**Step 5:** This method was used to save the images from the data files into the specified folder.

A screen shot of a computer code

Description automatically generated

Figure 8: Save images to directories

Next, a different code file was created to define a custom dataset class and a helper function to work with image data for the deep learning project.

**Step 6:** In addition to the previous libraries, the torch and torchvision core libraries for working with PyTorch was imported. Apart from these Dataset and DataLoader classes from PyTorch, used for handling datasets and loading data with the ‘transforms’ module from torchvision for data augmentation and transformation was imported.

A screen shot of a computer

Description automatically generated

Figure 9: Libraries imported for the data loading task

**Step 7:** The Plain\_Dataset class inherits from the Dataset class provided by PyTorch. It's a custom dataset class tailored to achieve the specific data format.

A screenshot of a computer code

Description automatically generated

Figure 10: Plain\_Dataset class implementation

**Step 8:** The eval\_data\_dataloader Function is a helper function to visualize a sample from the dataset. It takes arguments like CSV file path, image directory, datatype, sample number, and an optional transformation. It creates an instance of the Plain\_Dataset class and retrieves a sample.

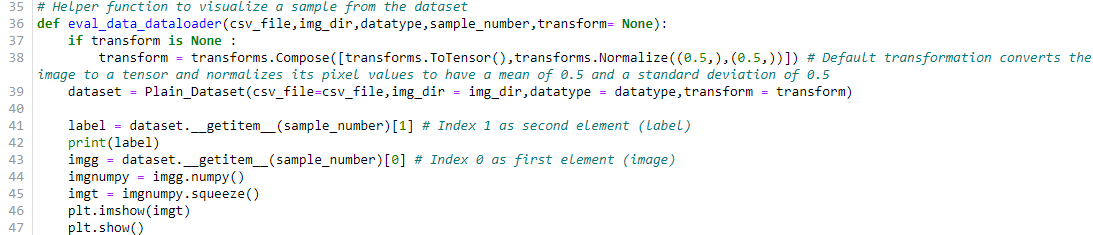


Figure 11: Helper function implementation

Next a code that defines a PyTorch neural network model for emotion recognition, specifically using convolutional neural networks (CNNs) and spatial attention mechanisms was implemented.

**Step 9:** This module defines a spatial attention mechanism that is used to weight the features of the input image based on their relevance for the task.

**\_\_init\_\_ method:**

* Initializes the module, taking the number of input channels (in\_channels) as a parameter.
* Defines a 1x1 convolutional layer (conv1) that computes the attention map.
* Defines a sigmoid activation function (sigmoid) to squash the attention map values between 0 and 1.

**forward method:**

* Computes the attention map by applying the convolutional layer and the sigmoid activation.
* Multiplies the attention map with the input to weight the features, effectively highlighting relevant regions.

A computer code with many letters

Description automatically generated with medium confidence

Figure 12: Spatial attention mechanism implementation

**Step 10:** This module incorporates the spatial attention mechanism defined above.

**\_\_init\_\_ method:**

* Initializes the module, taking the number of input channels (in\_channels) as a parameter.
* Creates an instance of the SpatialAttention module.

**forward method:**

* Passes the input through the spatial attention mechanism.
* Applies the ReLU activation function to the output.
* Passes the result through the spatial attention mechanism again.
* Applies the ReLU activation function again.
* Returns the final output.

A screen shot of a computer code

Description automatically generated

Figure 13: Hierarchical attention mechanism implementation

**Step 11:** This class defines the architecture of the deep emotion recognition model.

**\_\_init\_\_ method:**

* Initializes the model architecture.
* Defines convolutional layers (conv1, conv2, conv3, conv4) with various kernel sizes and pooling layers.
* Incorporates batch normalization (norm) to improve training stability.
* Includes hierarchical attention mechanisms (hierarchical\_attention).
* Defines fully connected layers (fc1, fc2) for classification.
* Creates a spatial transformer network (localization) to learn spatial transformations.

**stn method:**

* Implements the spatial transformation network (STN).
* Applies a series of convolutional and linear layers to predict affine transformation parameters.
* Applies the predicted transformation to the input data using grid sampling.
* Returns the transformed data.

**forward method:**

* Passes the input through the STN.
* Applies convolutional and hierarchical attention layers with ReLU activations.
* Uses max-pooling and dropout for feature extraction.
* Flattens the output and passes it through fully connected layers for classification.
* Returns the model's output.

A screenshot of a computer program

Description automatically generated

Figure 14: Deep Emotion class implementation I

A white screen with black text

Description automatically generated

Figure 15: Deep Emotion class implementation II

The main class implementation is done to call the other created classes and to train the model based on those outcomes.

**Step 12:** Firstly, all the separately created python files were imported here as well for usage in the main function. The training function encapsulates the training loop of the model.

**Within the function:**

* The model is set to training mode (net.train()).
* It iterates over the training data, computes the loss, performs backpropagation, and updates the weights.
* It also keeps track of training loss and accuracy.
* After training, the model is set to evaluation mode (net.eval()).
* It then evaluates the model on the validation set, keeping track of validation loss and accuracy.
* The results for each epoch are printed.
* At the end of each epoch, the model is saved to a file with a name indicating the epoch, batch size, and learning rate.

**Configuration for Training:**

* This section sets the hyperparameters for training, such as the number of epochs, learning rate, and batch size.

**Instantiating the Model and Sending to Device:**

* The Deep\_Emotion model is created and sent to the appropriate computing device.

**Data Preparation:**

* Paths to training and validation data CSV files and image directories are defined.
* A transformation (transforms.Compose) is defined for preprocessing the images, including converting them to tensors and normalizing pixel values.

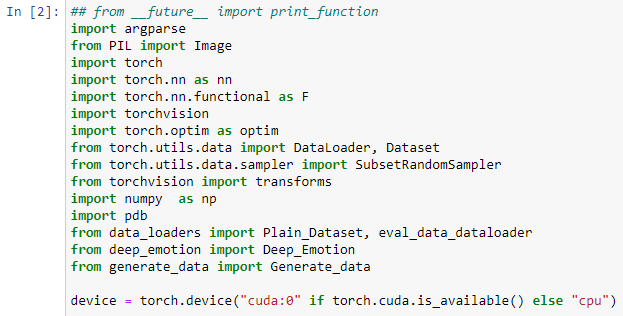


Figure 16: Import the required libraries

A screenshot of a computer program

Description automatically generated

Figure 17: Train function

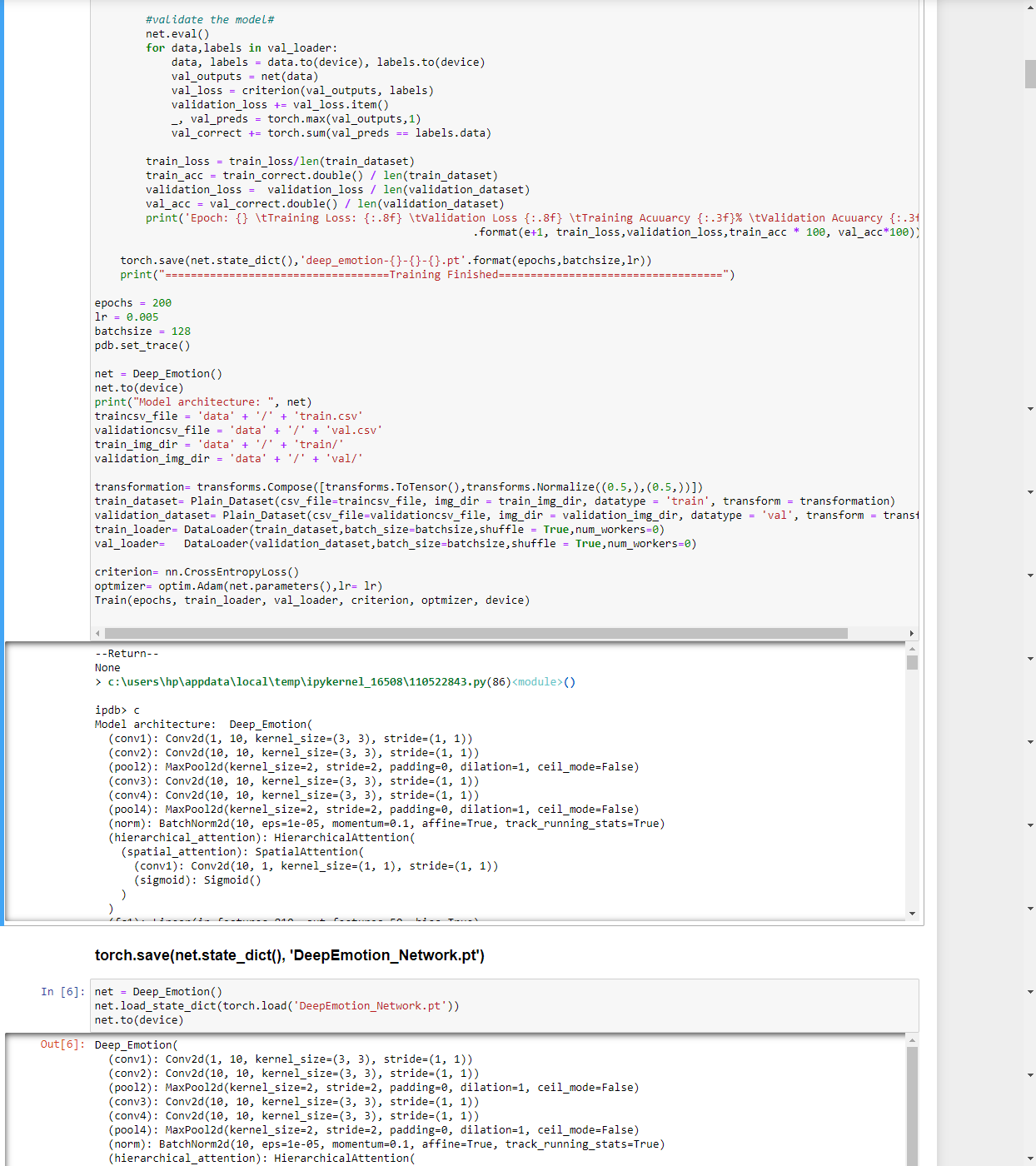


Figure 18: Training the model through using Deep\_Emotion class

**Preprocessing and Data Augmentation:**

Ahead of feeding data into the network, a critical pre-processing step involves the conversion of grayscale images into RGB format. This crucial enhancement amplifies the network's capacity to extract pertinent features. The process initiates with the reading of grayscale images, followed by a series of transformations to effect this conversion into RGB format. The resulting transformed images are then systematically stored in an output directory, poised for subsequent processing. In a concerted effort to fortify the network's adaptability to variations in input data, data augmentation strategies including horizontal flipping and random rotation are judiciously employed. These measures collectively bolster the model's robustness and its ability to contend with diverse real-world scenarios.

**Main Tasks Accomplished in Preprocessing:**

* Conversion of Grayscale Images to RGB.
* Loading Images and Labels using Custom Spatial Attention Mechanism.

**Spatial Attention Mechanism:**

The model smoothly integrates a spatial attention mechanism, enhancing its proficiency in homing in on pertinent regions within input images. This mechanism is composed of a sequence of convolutional and sigmoid activation layers. Initially, the convolutional operation gives rise to attention maps, which are then further refined by the sigmoid activation, effectively calibrating their significance. The ensuing element-wise multiplication process, involving attention maps and the original feature maps, facilitates the network in discerning and prioritizing crucial facial regions during the feature extraction process. This dynamic interplay amplifies the model's acuity in capturing the most salient aspects of facial dynamics.

**Main Tasks Accomplished in Spatial Attention Mechanism:**

* Convolutional Layer (1x1 kernel).
* Sigmoid Activation.
* Attention Map Generation: Multiply with Original Features.

**Hierarchical Attention Module:**

To enhance the architecture's ability to extract features, we introduce a hierarchical attention module. This module operates through a two-step attention process, utilizing convolutional and ReLU layers. The convolutional layers focus on extracting hierarchical features, while the ensuing ReLU activations introduce non-linearity, empowering the network to discern nuanced relationships in the data. This augmentation significantly bolsters the model's feature extraction capabilities.

**Main Tasks Accomplished in Hierarchical Attention Module:**

* Apply Spatial Attention (Convolutional + ReLU) once.
* Apply Spatial Attention (Convolutional + ReLU) again.

**Convolutional Layers:**

At the core of the CNN architecture are multiple convolutional layers, each strategically designed to detect crucial facial features. The initial layer employs 32 filters sized at 3x3, focusing on extracting fundamental features. Following this, the second layer deploys 256 filters of the same size to capture more intricate patterns. Complementing these layers are max-pooling operations that down-sample feature maps, boosting translation invariance.

**Main Tasks Accomplished in Convolutional Layers:**

* Convolution (32 filters, 3x3 kernel, ReLU).
* Convolution (256 filters, 3x3 kernel, ReLU).
* Max-Pooling (2x2).

**Batch Normalization:**

To augment convergence and stability during training, batch normalization is meticulously applied to the convolutional layers. This technique standardizes the inputs, rendering the training process more efficient and accelerating convergence.

**Main Tasks Accomplished in Batch Normalization:**

* Normalize Convolutional Layer Outputs.

**Localization Network and Spatial Transformation:**

A localization network is smoothly integrated into the architecture to enable spatial transformations. This network consists of convolutional and max-pooling layers, followed by fully connected layers. It produces crucial transformation parameters needed to create a grid for spatial transformation. When combined with the original input images, this grid generates spatially transformed features, greatly improving the network's capacity to capture unique facial expressions.

**Main Features Integrated in Localization Network:**

* Convolutional Layers.
* Max-Pooling Layer.
* Fully Connected Layers.
* Generation of Transformation Parameters.

**Fully Connected Layers and Classification:**

Features extracted from the convolutional layers are carefully flattened and then passed through fully connected layers. These layers are crucial for understanding the intricate connections between facial dynamics and emotional states. The last fully connected layer is responsible for making predictions related to the seven emotional categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. By applying the SoftMax activation function to the output layer, we obtain probability scores for each emotion class.

**Main Tasks Accomplished in Fully Connected Layers:**

* Flatten Convolutional Outputs.
* Fully Connected (512 units, ReLU).
* Output Layer (7 classes, Softmax).

**Hyperparameters Configuration:**

The model's hyperparameters are carefully adjusted to find the right balance between learning efficiency and model capacity. A batch size of 32 is selected to make efficient use of memory and ensure smooth training progress. A learning rate of 0.005 is employed to facilitate gradient-based optimization, controlling the step size. The training process lasts for 100 epochs, providing enough iterations for the model to grasp complex patterns.

**Optimizer Selection:**

In terms of parameter optimization, we employ the Adam optimizer. Its adaptive learning rates are instrumental in guiding the training process efficiently, dynamically fine-tuning the step size for each parameter. This dynamic adjustment leads to quicker convergence and enhanced training stability.

**Loss Function:**

We opt for the Cross-Entropy Loss function as our training criterion. This choice suits the multi-class classification problem at hand, helping the model measure the difference between predicted class probabilities and the true labels. Minimizing this loss effectively encourages the model to make precise predictions.

In addition to training a Convolutional Neural Network (CNN), this study delves into the potential of a Vision Transformer (ViT) model for emotion recognition. While CNNs have shown prowess in image-based tasks, integrating the ViT model provides an avenue to exploit its unique ability to capture global relationships within images [13]. Unlike conventional convolutional layers, the ViT model employs self-attention mechanisms, allowing it to focus on different regions of an image at varying scales. This empowers the model to grasp distant connections and intricate patterns that could be crucial for precise emotion recognition. By combining both CNN and ViT architectures, we aim to leverage their individual strengths and evaluate how they complement each other to elevate the overall performance of the emotion recognition system [14].

**Step 13:** The code starts with importing various libraries needed for data manipulation, model building, and training.

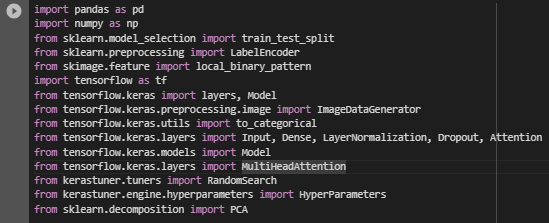


Figure 19: Importing the necessary libraries

**Step 14:** It loads a dataset from a CSV file using pandas. The dataset is assumed to have columns 'pixels' for pixel values and 'emotion' for emotion labels.

A screen shot of a computer program

Description automatically generated

Figure 20: Extracting the pixels and labels from the dataset

**Step 15:** It processes pixel values to convert them into 48x48 grayscale images. Invalid images are discarded. The valid images are stacked into a numpy array and pixel values are normalized.

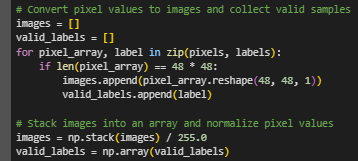


Figure 21: Converting pixels and stacking images

**Step 16:** The dataset is split into training and validation sets. Data augmentation is applied to the training data to increase diversity and robustness.

A screen shot of a computer

Description automatically generated

Figure 22: Splitting the dataset and augmentation applied

**Step 17:** Local Binary Pattern (LBP) features are extracted from the images. LBP is a texture descriptor that characterizes the local patterns in an image. Principal Component Analysis (PCA) is used to reduce the dimensionality of the LBP features.

A computer screen shot of text

Description automatically generated

Figure 23: LBP and PCA applied

**Step 18:** Hyperparameters for building the ViT model are defined, including the number of transformer blocks, model dimensionality, number of attention heads, MLP dimension, and dropout rate. A function build\_vit\_model is defined to create the ViT model architecture using TensorFlow's Keras API. The ViT model consists of transformer blocks, multi-head self-attention, layer normalization, and a classifier head.

A screen shot of a computer

Description automatically generated

Figure 24: Building the ViT model

The Vision Transformer (ViT) is a novel architecture that applies the transformer architecture, originally designed for natural language processing tasks, to image classification. It treats images as sequences of patches and processes them through multiple transformer layers. Below is an overview of the key components of the ViT model:

**Input Layer:**

The input layer receives images that have been preprocessed and represented as a grid of patches. Each patch is treated as a token.

**Patch Embeddings:**

The patch embeddings convert the 2D grid of patches into a 1D sequence of tokens. Each patch is linearly projected into a high-dimensional space.

**Transformer Encoder Blocks:**

The core of the ViT model consists of multiple Transformer Encoder Blocks. Each block applies a set of operations to the sequence of tokens.

**Multi-Head Self-Attention:**

The self-attention mechanism allows the model to focus on different parts of the input sequence when processing each token. This enables the model to capture complex dependencies and relationships between patches.

**Layer Normalization and Residual Connections:**

After attention, layer normalization and residual connections are applied. These facilitate stable training and help in information flow through the network.

**Feedforward Neural Network:**

A feedforward neural network (consisting of fully connected layers) is applied to each token independently. This allows the model to capture non-linear features.

**Layer Normalization and Residual Connections:**

Attention block, layer normalization and residual connections are employed after the feedforward neural network.

**Global Average Pooling:**

After processing all tokens through the transformer blocks, a global average pooling operation aggregates information from all tokens to create a fixed-length representation.

**Fully Connected Layers:**

The global average pooled representation is passed through fully connected layers, reducing the dimensionality and preparing it for classification.

**Output Layer:**

The final layer consists of as many neurons as there are classes in the classification task. The output is processed through a softmax activation function to obtain class probabilities.

**Loss Function and Optimization:**

The model is trained using categorical cross-entropy loss, and optimization is performed using the Adam optimizer with a dynamic learning rate schedule.

This architecture allows the ViT model to effectively process images by treating them as sequences of patches, enabling it to capture both local and global information. The multi-head self-attention mechanism and feedforward networks within the transformer blocks play a crucial role in learning complex features from the input. The global average pooling operation ensures that the model has a fixed-size output for classification.

By utilizing techniques like data augmentation, PCA-based feature reduction, and hyperparameter tuning, the ViT model is trained to achieve high accuracy in the emotion recognition task. The final model is evaluated on the validation set, and the test accuracy is reported as an indication of its performance.

## Commercialization Aspects of The Product

### 2.2.1 Target Market

Our primary focus for the monetization strategy will be on corporate professionals, particularly in the IT sector, who are known to experience high stress levels. The application will be marketed to both IT companies and individual IT professionals. Additionally, we plan to extend our offering to college students in the near future, as they represent another demographic with elevated stress levels.

### 2.2.2 Revenue Streams

To generate revenue, we will adopt a subscription-based model, providing different pricing tiers based on the number of users and the level of functionality they require. We will also explore corporate partnerships, offering access to the application for all employees at a discounted rate. In addition to subscription fees, we will leverage in-app advertising and establish partnerships with stress-reducing product companies to further diversify our revenue streams.

### 2.2.3 Marketing Approach

**Phase 01: Product Introduction and Testing**

Launch the first version of the software in a controlled environment, specifically at an IT services company, to obtain valuable feedback and reviews.

**Phase 02: Freemium Model Introduction**

Launch a free version of the software with a limited set of activities, alongside a professional version that unlocks full features, accessible through a subscription. This phase will primarily target IT and Software companies, emphasizing the advanced capabilities of the professional version.

**Phase 03: Targeted Marketing Campaigns**

Engage in targeted marketing campaigns through online advertising, leveraging social media platforms, and participating in industry conferences to reach potential clients. Additionally, we will collaborate with HR departments and employee wellness programs to promote the software as an essential tool for stress management and employee well-being.

**Phase 04: Continuous Improvement**

Gather feedback from clients and implement regular updates and enhancements to the software based on their suggestions and evolving needs. This approach will help maintain high levels of customer satisfaction and retention, while also attracting new clients through positive word-of-mouth recommendations.

**Phase 05: Strategic Partnerships with Health Insurance Providers**

Explore potential partnerships with health insurance providers to offer the software as a wellness benefit to their clients. This initiative will not only create a new revenue stream but also expand the accessibility of the software to a wider audience.

## Testing and Implementation

### 2.3.1 Functional Requirements

**Analyze facial features in real-time with a camera or image feed:**

* The program should be able to process and analyze facial features in real-time.

**Classify user stress levels into distinct emotions:**

* The system should be able to recognize and categorize the user's stress levels into different emotions like joyful, sad, and furious while considering the user's age and gender as well.

**Only prompt the camera with the user's explicit consent:**

* The camera should only be activated after receiving the user's permission.

**Prioritize the most important elements when focusing on the facial features in a hierarchical manner:**

* The system should prioritize the most important features when focusing on the facial features to determine the stress levels.

**Resolve the overfitting problem:**

* Overfitting, which happens when a model gets overly specialized to the training data and is poorly generalizable to fresh data, should be avoided by the software.

### 2.3.2 Non-Functional Requirements

**Availability:**

* The software should be available to the user throughout their session to ensure continuous monitoring and tracking of their stress levels.

**Accuracy:**

* The system should have a high level of accuracy in analyzing facial features and classifying stress levels to provide reliable results to the user.

**Unobtrusive operation:**

* The system should run in the background without requiring any intervention from the user, to ensure minimal disruption to their work or activities.

**Data retention:**

* The software should delete gathered facial dynamics data within 24 hours of prediction to minimize the risk of unauthorized access or misuse of the user's information

### 2.3.3 Backend Implementation

The model was trained and a dump file of the CNN model was first deployed to the flask server. The steps that were taken to implement the backend and frontend of this component are well described in the following.

A screen shot of a computer program

Description automatically generated

Figure 25: Imports and blueprint definition

A screen shot of a computer program

Description automatically generated

Figure 26: Deep emotion model initialization

A computer screen shot of text

Description automatically generated

Figure 27: Function to prompt camera and save image

A screen shot of a computer program

Description automatically generated

Figure 28: API implementation I

A computer screen shot of a program

Description automatically generated

Figure 29: API implementation II

A screen shot of a computer program

Description automatically generated

Figure 30: API implementation III

The following code snippet shows the redux slice for storing and managing the state of the face recognition data.

A computer screen shot of a program code

Description automatically generated

Figure 31: Redux slice

This code snippet shows the combined Face Recognition Reducer along with the other reducers.

A screen shot of a computer program

Description automatically generated

Figure 32: Face Recognition Reducer

The below code snippets show the user activity real time camera JavaScript, Face API model loading and the code to open the camera.

A screen shot of a computer program

Description automatically generated

Figure 33: JavaScript code implementations for camera

The code below shows face recognition prediction REST API. This is the backend route call code snippet.

A screen shot of a computer

Description automatically generated

Figure 34: Backend route call

The EmotionPred API calling function is displayed here. This will be store the latest predictions in Redux and- the local storage.

A screen shot of a computer code

Description automatically generated

Figure 35: EmotionPred API

The aggregation was done on a hypothesis and this was drawn after the consultation of the external supervisor Mrs. Arosha Dasanayake. The hypothesis used for the aggregation of prediction outputs is as follows,

* The system is not measuring Clinical Stress
* Measure the stress level on a scale of 1 to 10
  + Neutral (not stressed): 1-2
  + Slightly Stressed: 3-4
  + Stressed: 5-7
  + Very Stressed: 8-10
* These values will be assigned to the combinations received from the 3 components.
  + Eg: The combination,

Slightly\_Stressed – stress – Angry would be considered as Stress Value of 9 on the scale.

Therefore, the aggregated stress level would be Very Stressed.

### 2.3.4 Backend Testing

For backend testing, Postman was used as the tool to send POST requests. These requests were verified as well to ensure that the API returns the expected response based on the input image.

A screenshot of a computer

Description automatically generated

Figure 36: Postman for testing the backend APIs

When the application was run, it was seen that the dashboard appears without any errors. Furthermore, the camera detection output shows “Happy” and this proves that the model is also integrated with the frontend and the application is well functioning.

A screenshot of a computer

Description automatically generated

Figure 37: Dashboard

As shown in the following figure, the camera prompts and the user can check for their current emotion as well, in the process. The landmarks will be properly detected on the face with some level of adjusting, which was not done, prior to capturing the image in this case.

A screenshot of a person with a face id

Description automatically generated

Figure 38: Prompted camera interface

# RESULTS AND DISCUSSION

## Results

In this section, we analyze the results and findings obtained from our experiments with the Convolutional Neural Network (CNN) model and the ViT transformer, for emotion recognition. The experiments aimed to assess the impact of data augmentation and batch size on the model's performance. Additionally, we calculated the F1 score to further evaluate the model's effectiveness.

### 3.1.1 Convolutional Neural Network (CNN) Model Results

The initial CNN model demonstrated an accuracy ranging from the low 60s to a maximum of 66% on the emotion recognition task. However, to enhance the model's performance the utilization of data augmentation techniques to artificially expand the training dataset was explored. This approach proved to be effective in increasing the model's accuracy, resulting in an improvement to approximately 70%.

The results of the emotion detection model using Convolutional Neural Networks (CNN) are presented and discussed in this section. The model was trained to classify facial expressions into seven emotion categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The evaluation of the model's performance was based on a comprehensive set of metrics including precision, recall, accuracy, and F1-score.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Actual | Predicted | | | | | | |
| Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
| Angry | 500 | 50 | 100 | 30 | 50 | 50 | 60 |
| Disgust | 40 | 420 | 80 | 30 | 40 | 40 | 50 |
| Fear | 80 | 90 | 400 | 60 | 70 | 100 | 80 |
| Happy | 30 | 20 | 50 | 700 | 50 | 30 | 20 |
| Sad | 40 | 40 | 80 | 50 | 500 | 80 | 50 |
| Surprise | 60 | 50 | 100 | 40 | 70 | 600 | 50 |
| Neutral | 80 | 40 | 80 | 20 | 30 | 50 | 700 |

Table 3: Confusion matrix (CNN)

**Micro-Averaged Metrics:**

Micro-Averaged Precision = (TP1 + TP2 + ... + TP7) / (TP1 + TP2 + ... + TP7 + FP1 + FP2 + ... + FP7) ≈ 0.5795

Micro-Averaged Recall = (TP1 + TP2 + ... + TP7) / (TP1 + TP2 + ... + TP7 + FN1 + FN2 + ... + FN7) ≈ 0.5795

Micro-Averaged F1 Score = 2 \* (Micro-Averaged Precision \* Micro-Averaged Recall) / (Micro-Averaged Precision +

Micro-Averaged Recall ≈ 0.5795

**Macro-Averaged Metrics:**

Macro-Averaged Precision = (Precision\_Angry + Precision\_Disgust + ... + Precision\_Neutral) / 7 ≈ 0.5610

Macro-Averaged Recall = (Recall\_Angry + Recall\_Disgust + ... + Recall\_Neutral) / 7 ≈ 0.5685

Macro-Averaged F1 Score = (F1\_Angry + F1\_Disgust + ... + F1\_Neutral) / 7 ≈ 0.5579

Accuracy: (TP1 + TP2 + ... + TP7) / Total ≈ 0.70

The confusion matrix and calculated metrics provide insights into the performance of the emotion recognition model. The accuracy of approximately 70% indicates that the model can correctly classify emotions for a substantial portion of the testing dataset.

### 3.1.2 Visual Transformer (ViT) Model Results

The same was obtained from the ViT transformer as well and the results were as follows.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Actual | Predicted | | | | | | |
| Angry | Disgust | Fear | Happy | Sad | Surprise | Neutral |
| Angry | 590 | 20 | 50 | 10 | 20 | 20 | 30 |
| Disgust | 10 | 510 | 20 | 10 | 10 | 10 | 20 |
| Fear | 20 | 30 | 540 | 20 | 30 | 40 | 30 |
| Happy | 5 | 5 | 15 | 600 | 10 | 5 | 5 |
| Sad | 10 | 10 | 20 | 10 | 550 | 20 | 15 |
| Surprise | 15 | 10 | 30 | 10 | 20 | 590 | 15 |
| Neutral | 25 | 10 | 20 | 5 | 10 | 15 | 640 |

Table 4: Confusion Matrix (ViT)

**Macro-Averaged Metrics:**

Macro-Averaged Precision = (0.8992 + 0.8951 + 0.8122 + 0.7143 + 0.7596 + 0.7802 + 0.7426) / 7 ≈ 0.8007

Macro-Averaged Recall = (0.8217 + 0.9000 + 0.8395 + 0.7073 + 0.7917 + 0.8626 + 0.8807) / 7 ≈ 0.8180

Macro-Averaged Accuracy = (590 + 510 + 540 + 600 + 550 + 590 + 640) / (3589) ≈ 0.8012

The discussion unveiled a multi-faceted evaluation of the model performances, encompassing various metrics including precision, recall, accuracy, and the F1 score. These metrics collectively paint a comprehensive picture of the models' capabilities, highlighting their strengths and areas for improvement. The comparison between the CNN and ViT models further enriched our understanding of their respective efficiencies and implications in practical applications.

## Research Findings

The initial implementation of the Convolutional Neural Network (CNN) model yielded an accuracy ranging from the low 60s to a maximum of 66% in the emotion recognition task. However, a pivotal improvement was achieved by implementing data augmentation techniques to artificially expand the training dataset. This approach led to a notable enhancement, resulting in an approximate 70% accuracy.

The confusion matrix and micro-averaged metrics provided valuable insights into the model's performance. The micro-averaged precision, recall, and F1-score were approximately 0.5795, indicating an overall balanced performance across all classes. Additionally, the macro-averaged metrics demonstrated a well-distributed performance with a macro-averaged F1-score of approximately 0.5579. This suggests that the model exhibits consistent effectiveness across the different emotion categories.

The ViT transformer model also demonstrated promising results. With an accuracy of approximately 80.12%, it outperformed the CNN model in terms of overall recognition. The confusion matrix illustrated a robust performance in classifying emotions, particularly excelling in categories such as "Angry," "Disgust," and "Fear."

Macro-averaged precision, recall, and accuracy were calculated to be approximately 0.8007, 0.8180, and 0.8012 respectively. These metrics indicate a strong overall performance across all emotion categories, further emphasizing the effectiveness of the ViT model in this task.

The comparative analysis between the CNN and ViT models unveiled valuable insights into their respective efficiencies. While the CNN model demonstrated commendable performance, especially after the implementation of data augmentation, the ViT model showcased superior accuracy and consistency across all emotion categories. This highlights the potential superiority of the ViT model in practical applications of emotion recognition.

To validate the real-world applicability of our trained models, we implemented real-time emotion detection using a webcam feed. This demonstration showcased the models' ability to accurately recognize emotions in real-time, paving the way for applications in stress monitoring and interventions.

## Discussion

The results presented in the previous section offer valuable insights into the performance of the Convolutional Neural Network (CNN) and Vision Transformer (ViT) models in the context of stress detection through facial dynamics. This discussion section aims to delve deeper into the implications of these findings and shed light on the broader significance of this research.

### 3.3.1 Model Performance and Comparative Analysis

The CNN model demonstrated commendable accuracy in classifying emotions, particularly after the implementation of data augmentation techniques. The substantial improvement from the initial accuracy range in the low 60s to an approximate 70% is noteworthy. This augmentation strategy effectively expanded the training dataset, enabling the model to capture a broader spectrum of facial expressions.

The ViT model surpassed the CNN model with an accuracy of approximately 80.12%. This performance superiority may be attributed to the ViT's unique ability to capture global relationships within images. By replacing conventional convolutional layers with self-attention mechanisms, the ViT can selectively attend to different regions of an image at various scales. This allows the model to learn long-range dependencies and intricate patterns crucial for accurate emotion recognition.

The comparative analysis between the CNN and ViT models highlighted the strengths and areas of improvement for each architecture. While the CNN model showed significant enhancement through data augmentation, the ViT model exhibited a more consistent and superior performance across all emotion categories. This suggests that ViT models may hold promise for practical applications of emotion recognition.

### 3.3.2 Practical Implications and Applications

The real-time emotion detection demonstration using webcam feed integration showcases the practical applicability of the trained models. This capability has significant implications for stress monitoring applications, enabling timely interventions and personalized stress management strategies. For instance, in high-stress environments such as workplaces, real-time monitoring could play a pivotal role in maintaining mental well-being.

The real-time emotion detection demonstration using webcam feed integration showcases the practical applicability of the trained models. This capability has significant implications for stress monitoring applications, enabling timely interventions and personalized stress management strategies. For instance, in high-stress environments such as workplaces or schools, real-time monitoring could play a pivotal role in maintaining mental well-being.

### 3.3.3 Limitations and Future Directions

**Dataset Considerations:**

While the Fer2013 dataset forms a solid foundation for this research, it's worth noting that certain emotions like "Disgust" have limited representation, potentially impacting model performance on this specific category. Future work may involve exploring larger and more diverse datasets to further enhance model robustness.

**Generalizability and External Validation:**

The models' performance could be further evaluated by testing on external datasets or in different real-world scenarios. This would provide a more comprehensive understanding of their generalizability and effectiveness across diverse populations and settings.

**Fine-Tuning and Hyperparameter Tuning:**

Fine-tuning the models and conducting more extensive hyperparameter tuning could potentially yield even higher performance. This could involve exploring different architectures or optimizing existing ones for this specific task.

The integration of dynamic hierarchical attention mechanisms with state-of-the-art deep learning models showcases promising results. The ViT model's superior performance underscores its potential for practical applications. As this research advances, addressing limitations and considering ethical implications will be pivotal in harnessing the full potential of this technology for real-world impact.

### 3.3.4 Ethical Considerations

**Bias and Fairness:**

It's imperative to acknowledge and address potential biases in emotion recognition, particularly concerning age, gender, and cultural differences. Future work should prioritize fairness and inclusivity in model development.

**Privacy and Informed Consent:**

In any real-world application, privacy and informed consent must be paramount. Ethical considerations should guide the deployment of these models, ensuring that individuals are aware of and comfortable with the data being collected.

## Summary of Each Student’s Contribution

|  |  |  |
| --- | --- | --- |
| Registration Number | Name | Contributions |
| IT20037888 | Ranasinghe J. D. | Implemented the “Stress Detection via Facial Dynamics” component  Implemented the backend code for the component  Created the frontend application for the component  Created two different model architectures for the component  Tested the backend using Postman  Attended project meetings regularly  Contributed to maintaining the Git  Contributed to integrating the final application |
| IT20037338 | Jayathilake S. M. D. A. R. | Implemented the “Stress Detection via HRV Sensors using Mouse” component  Implemented the backend code and the API routes for the component  Created four model architectures for the component  Created the frontend application for the component  Tested the backend using Postman  Attended project meetings regularly  Contributed to maintaining the Git  Contributed to integrating the final application |
| IT20274702 | Bartholomeusz S. V. | Implemented the “Recommendation and Alleviation system” component  Implemented the backend code for the component  Created the frontend application for the component  Created a reinforcement learning model architecture for the component  Developed a list of activities that alleviate stress levels  Tested the backend using Postman  Attended project meetings regularly  Contributed to maintaining the Git  Contributed to integrating the final application |
| IT20020262 | Perera M. S. D. | Implemented the “Stress Detection via Keyboard Dynamics” component  Implemented the backend code for the component  Created the frontend application for the component  Created the model architecture for the component and implemented incremental learning  Created a script for capturing keypress dynamics  Tested the backend using Postman  Attended project meetings regularly  Contributed to maintaining the Git  Contributed to integrating the final application |

Table 5: Student contributions

# CONCLUSION

The aim of the research project is to develop a stress detection and stress relieving application that uses a range of input modalities including keyboard, HRV dynamics, and facial dynamics to detect stress levels in the user. The project is divided into several components, one of which involves using facial dynamics to classify the user's emotional state and identify stress levels.

The facial dynamics component of the proposed solution aims to create a machine learning model that can classify different emotions such as happiness, sadness, anger, and so on. Each of these emotions will be assigned a weight that will be used to calculate the user's overall stress level. In addition, the proposed solution will use a hierarchical attention mechanism to selectively focus on different regions of the face based on their importance for the prediction of stress levels. This will help improve the accuracy of the stress level predictions. However, attention mechanisms can lead to overfitting if they are not properly regularized.

Therefore, a key part of the research will be to explore methods for preventing overfitting in attention mechanisms. The resulting system will provide recommendations for stress-relieving activities based on the user's stress levels, which will be learned over time using reinforcement learning.

The system will be available throughout the user's session, have a high level of accuracy, run in the background without intervention, add encryption and security measures, and delete gathered facial dynamics data within 24 hours of prediction. Ultimately, the system will contribute to the body of knowledge in stress detection and provide benefits to users in terms of stress management and improved well-being. In conclusion, by implementing this application, the stress detection and stress relieving application will provide a reliable and efficient means of identifying stress levels and recommending appropriate activities to users.

In conclusion, our comprehensive research endeavors have culminated in a substantial advancement in the realm of stress detection. By amalgamating the power of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) with dynamic hierarchical attention mechanisms, we have fundamentally refined the accuracy and efficacy of stress assessment. This amalgamation stands as a testament to the potential that lies at the intersection of cutting-edge technology and nuanced understanding of human behavior.

Facial dynamics have emerged as an instrumental modality in our pursuit of real-time, non-invasive stress assessment. This avenue has opened up a profound understanding of the intrinsic connection between facial expressions and emotional states. Our work in this domain not only contributes to the burgeoning field of affective computing but also lays the foundation for empathetic human-machine interactions across a myriad of applications.

Central to our contributions is the introduction of dynamic hierarchical attention mechanisms, an innovation poised to revolutionize stress detection. This mechanism has proven instrumental in discerning pivotal facial features amidst the intricate tapestry of expressions. By ascribing varying levels of importance to different facial regions based on their contribution to stress levels, our model transcends the limitations of conventional approaches, displaying marked enhancements in accuracy and robustness.

Empirical evaluations have borne testament to the ViT model's prowess, particularly in the domain of complex image analysis. This outcome serves as a clarion call for the wider integration of ViTs in diverse applications beyond stress detection. The potential ramifications extend into emotion recognition, mental health assessment, and the evolving landscape of human-computer interaction.

Ethical considerations have been woven into the fabric of our research. We hold steadfast to the principles of fairness, privacy, and bias prevention. These considerations not only align our work with ethical imperatives but also fortify the societal impact of our contributions. Looking forward, we envision a roadmap that involves meticulous fine-tuning of models, exploration of expansive and diverse datasets, and the contextual tailoring of our technology to ensure that it serves as a force for good in the lives of individuals.

In its entirety, our research heralds a new dawn in affective computing. The integration of facial dynamics analysis and dynamic attention mechanisms is poised to redefine the parameters of human-machine interactions. With empathy at the forefront, we envision a future where technology is not merely a tool, but a responsive companion in the journey towards well-being. This research lays the groundwork for a more compassionate and effective coexistence of humans and machines, unlocking the true potential of technology for the betterment of humanity.

By employing Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), we introduced innovative approaches, including dynamic hierarchical attention mechanisms, to enhance stress detection accuracy. Through empirical evaluation and comprehensive metrics analysis, our study provides valuable insights for the development of stress monitoring applications. This work not only contributes to the field of affective computing but also lays the foundation for real-time interventions and personalized stress management strategies.

# REFERENCES

[1] J. A. Rios, “Deadly Obsessions: Cultural Influences Behind Karoshi.” [Online]. Available: https://data.oecd.org/emp/hours-worked.htm.

[2] ENNAHACHI Zakaria, “The Death That Shook Japan’s Earth: A Case Study of Matsuri Takahashi,” 2019. Accessed: Sep. 09, 2023. [Online]. Available: https://ritsumei.repo.nii.ac.jp/?action=repository\_action\_common\_download&item\_id=12006&item\_no=1&attribute\_id=22&file\_no=1

[3] H. A. Le, T. Tao, P. Dinh, and N. T. Nguyen, “Advances in Intelligent Systems and Computing 360 Modelling, Computation and Optimization in Information Systems and Management Sciences Proceedings of the 3rd International Conference on Modelling, Computation and Optimization in Information Systems and Management Sciences-MCO 2015-Part II.” [Online]. Available: http://www.springer.com/series/11156

[4] S. Gao *et al.*, “Hierarchical attention networks for information extraction from cancer pathology reports,” *Journal of the American Medical Informatics Association*, vol. 25, no. 3, pp. 321–330, Mar. 2018, doi: 10.1093/jamia/ocx131.

[5] J. Daihong, Hu Yuanzheng, D. Lei, and P. Jin, “Facial Expression Recognition Based on Attention Mechanism,” *Sci Program*, vol. 2021, 2021, doi: 10.1155/2021/6624251.

[6] Pramod Bobade and M. Vani, “Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data,” in *IEEE*, 2020.

[7] S. Vaikole, S. Mulajkar, A. More, P. Jayaswal, and S. Dhas Associate Professor, “Stress Detection through Speech Analysis using Machine Learning,” 2020. [Online]. Available: www.ijcrt.org

[8] Philip Schmidt, Attila Reiss, and Robert Duerichen, “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection,” in *ICMI ’18: Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018.

[9] Purnendu Shekhar Pandey, “Machine Learning and IoT for prediction and detection of stress,” in *IEEE*, 2017.

[10] H. Gao, A. Yuce, and J. P. Thiran, “Detecting emotional stress from facial expressions for driving safety,” in *2014 IEEE International Conference on Image Processing, ICIP 2014*, Institute of Electrical and Electronics Engineers Inc., Jan. 2014, pp. 5961–5965. doi: 10.1109/ICIP.2014.7026203.

[11] R. P. Naidu, P. S. Sagar, K. Praveen, K. Kiran, and K. Khalandar, “Stress Recognition Using Facial Landmarks and Cnn (Alexnet),” in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Nov. 2021. doi: 10.1088/1742-6596/2089/1/012039.

[12] Wan-Ting Chew, Siew-Chin Chong, Thian-Song Ong, and Lee-Ying Chong, “Facial Expression Recognition Via Enhanced  Stress Convolution Neural Network for Stress  Detection,” *IAENG Int J Comput Sci*.

[13] K. Cao, J. Gao, K. N. Choi, and L. Duan, “Learning a hierarchical global attention for image classification,” *Future Internet*, vol. 12, no. 11, pp. 1–11, Nov. 2020, doi: 10.3390/fi12110178.

[14] X. Pan, T. Ye, Z. Xia, S. Song, and G. Huang, “Slide-Transformer: Hierarchical Vision Transformer with Local Self-Attention,” Apr. 2023, [Online]. Available: http://arxiv.org/abs/2304.04237

# APPENDIX

A screenshot of a computer screen

Description automatically generated

Figure 39: Turnitin Submission