INTERNSHIP/ DISSERTATION I REPORT

ON

REAL-TIME ADVANCED MULTI-FACE EMOTION RECOGNITION AND SENTIMENT ANALYSIS USING KAFKA AND VIT

Submitted in partial fulfilment of the requirements for the degree of

M.Tech Computer Science and Engineering

(Big Data Analytics)

by

YESU RAGUL J

24MCB0041

Under the Supervision of

N. Harini

Assistant Professor

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

November, 2025

DECLARATION

I, **Yesu Ragul J (24MCB0041)**, hereby declare that the capstone project titled "**Real-Time Advanced Multi Face Emotion Recognition and Sentiment Analysis using Kafka and ViT**" submitted to Vellore Institute of Technology (VIT), Vellore, for the conferment of the Master of Technology degree in Big Data Analytics, constitutes a genuine record of work conducted by me under the supervision of Assistant Prof. N. Harini , School of Computer Science and Engineering, Vellore Institute of Technology, Vellore.

I thus affirm that the work presented in this project report has not been submitted, nor will it be submitted, in whole or in part, for the conferment of any other degree or certificate at this institution or any other institution or university.

Place: Vellore Signature of the Candidate

Date:

CERTIFICATE

This is to certify that the capstone project entitled **"Real-Time Advanced MultiFace Emotion Recognition and Sentiment Analysis using Kafka and ViT"** submitted by **Yesu Ragul J (24MCB0041)**, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore for the award of the degree of **Master of Technology (Big Data Analytics)**, is a record of bonafide work carried out by him/her under my supervision during the period, **13-07-2025 to 17-11-2025** as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other Institute or University. The dissertation fulfills the requirements and regulations of the Institute and in my opinion meets the necessary standards for submission.

Place: Vellore, Name and signature of the guide

Date:

Internal Examiner **External Examiner**

Head of the Department

Department of HoD Department

ABSTRACT

Comprehending human emotions in real-time is increasing in significance for applications ranging across social analytics and assistive technologies. This dissertation presents a comprehensive real-time pipeline for emotion detection using Apache Kafka and a Vision Transformer (ViT) to identify and relay diverse facial expressions. Our system records facial images from a camera or dataset, preprocesses the images (recognition, cropping, resizing, and normalization), and sends the images along to a specifically trained ViT model for seven fundamental emotions. Then a consumer visualizes and/or logs outputs after the Kafka producer sends the images to a worker process utilizing the ViT to infer emotion and assign confidence scores. A Streamlit dashboard presents an interactive user interface, and all components are package encapsulated to enable seamless deployment. The system retains low latency for inference and maintains substandard classification performance while implementing a decentralized pipeline. Performance analyses using widely accessible public datasets show that our implementation is capable of consistently achieving over 90% classification accuracy and maintains latency of under 100 ms per frame of the new consumer application, showing modern transformer architectures and distributed streaming can implement fast processing and scalable real- time multi-face emotion detection for real-world implementation applications.

Keywords: *Vision Transformer (ViT), Real-Time Emotion Recognition, Multi-Face Detection, Sentiment Analysis, Apache Kafka, Facial Expression Recognition, Streaming Architecture, Deep Learning.*

ACKNOWLEDGEMENTS

With immense pleasure and a deep sense of gratitude, I wish to express my sincere thanks to my supervisor **Prof N. Harini**, School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore without his/her motivation and continuous encouragement, this research would not have been successfully completed.

I am grateful to the Chancellor of VIT, **Dr. G. Viswanathan**, the Vice Presidents, and the Vice Chancellor for motivating me to carry out research in the Vellore Institute of Technology.

It would be no exaggeration to say that Dean in-charge of SCOPE, **Prof. Jaisankar N,** was always available to clarify any queries and clear the doubts I had during the course of my project. I would also like to acknowledge the role of **HoD, Dr. Sendhil Kumar**, who was instrumental in keeping me updated with all necessary formalities and posting all the required formats and document templates through the mail, which I was glad to have had.

Lastly, I would like to thank the Vellore Institute of Technology to provide the facilities of the requisite infrastructure, provide flexibility of choice, and encourage my research and implementation of the work related to the dissertation.

Place: Vellore, **Yesu Ragul J (24MCB0041)**

Date:

CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Overview

Humans cannot be communicated without their emotions because they determine how individuals express their states of affection, react to the stimuli in the environment, and interact with computer interfaces. With the growing penetration of digital systems in daily activities, the ability of robotic agents to detect signs of an emotional state has become central to the success of human-computer communication. Expressions are one of the most powerful and universal measures of emotional states, which makes facial expression recognition (FER) one of the well-spread methods of evoking such feelings. Facial expressions are a strong and universal display of the emotional expressions therefore, the concept of facial expression recognition (FER) became a standard means of examining the signals of emotions. Older FER systems used features (such as Support Vector Machines (SVMs), Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP)) that had to be engineered manually. Although these techniques worked well in the controlled laboratory setting, they were faced with significant challenges when used in real world situations because of variations in illumination, occlusions which block visual accessibility and different face orientations. The introduction of deep-learning algorithms, especially Convolutional Neural Networks (CNNs) was a major improvement since it allowed extracting more complex facial patterns to recognize the emotional state. However, CNNs have restricted receptive fields that limit the receptivity of the model in grasping global spatial relations that would be critical in detecting subtle differences in expressing emotions. To mitigate this shortcoming, the Vision Transformer (ViT) model is provided with self-attention keys, which are applied to the whole image, which is why the network evaluates contextual interactions of distant regions of the face regardless of the spatial distance. The facial interactions accentuating across the world contribute to the development of a more in-depth depiction of emotional cues. This project creates a system, which takes image frames, analyses the affective states, and publishes the results without significant hold-up using ViT and Kafka. The very nature of Kafka as the capacity to handle high-frequency data streams, including such aspects as stream segmentation, fault resilience, and scalability horizontally, makes it a very appropriate candidate in terms of maintaining nonstop and real-time support in the given framework. The suggested architecture uses a Kafka Producer to materialize and introduce facial photographs out of a live camera perspective or offline information on separate Kafka subjects. ViT-driven Emotion Detection Worker then processes these images and consequently performs the preprocessing task and then the task of labeling its findings before piping the predictions into a Kafka consumer. Its methodology allows simultaneous examination of several faces on one frame

therefore allowing multi-faceted emotion detection. Overall, this project provides integrated sentiment analysis in the form of an integrated platform that combines deep-learning methods and big-data streaming setups. The resulting system can be used in a variety of fields, such as human-robot interaction, intelligent learning environments, security operations, entertainment analytics, mental health assessment, and customer experience research.

1.2 Objectives

The ultimate aim of the dissertation is to work on and realize a face emotion recognition system with the highest level of accuracy in real-time and scale. The proposed project aims at building a Vision Transformer (ViT) model that would be reliable to detect the seven basic human emotions, namely happiness, sadness, anger, surprise, fear, disgust, and neutrality. In order to get data processed in real time with minimum latency and maximum throughput to process emotion, the system will combine ViT framework with Apache Kafka. In addition, it will be furnished with functions of face detection capabilities in individual frame and determination of the subsequent emotions.

The architecture will be designed in a modular manner, which will ease interaction between data producers, the stakeholders, and consumers of the information. Precision, latency, capacity and resilience in real time scenarios will also be favored during performance evaluation. Moreover, the solution will have optimized APIs and containerization technologies to be deployed in distributed settings. Finally, it will use dashboards or formatted outputs to provide easy visualizations and analytics regarding emotion foretellings.

1.3 Motivation

With the rising interactions of AI systems with human beings in sensitive and dynamic environments, emotions awareness in real-time is an imperative requirement. The knowledge of affective states is paramount to improving the user experience and perfecting decision-making. Traditional facial emotion recognition systems are able to identify simple expressions but often do not work when there is an uncertain or unstable background in the area. The main weakness they have is that it is unable to capture relational associations between features of the face and this makes their application unreliable in practice. To solve this weakness, Vision Transformers (ViT) work with long-range long-range dependencies and captures fine-scale patterns in the area as a whole, which allow recognizing emotions in a more precise and reliable way under varying conditions. The Apache Kafka is an excellent option due to its high-throughput, fault tolerance, and good support of asynchronous communication in a distributed architecture. With Kafka and the power of Vision

Transformers, it is possible to build a complex and scalable image of the real-time face emotion recognition infrastructure. Such synergy allows quicker and more accurate affective understanding thus fulfilling the ever-growing need of smart AI systems capable of responding to human feelings.

1.4 Background

During the last ten years facial expression recognition has developed tremendously. Past methods have mostly used appearance-based or geometrical features. These models performed adequately in an ideal laboratory setting, but they did not successfully generalize to non-laboratory settings. With the development of machine learning methods that utilize deep learning, the ability to generalize has increased, but these models still struggle to cover the features that are considered worldly, to be able to differentiate subtle differences in emotions. Vision Transformers provided a new paradigm, which means the rejection of the traditional convolutional architecture altogether and the division of images into consecutive patches. The multi-head self-attention feature embedded in Vision Transformers can not only capture the local image detail, fiscal focus on the global contextual information, thus, it has been able to reach state-of-the-art performance in facial emotion recognition tasks.

The literature has empirical accuracy measures above 90 % on mainstream datasets, FER2013 and RAF -DB, extreme models can be more stable and less interpretable compared to dissatisfied CNN -based ones.The advantages of deep learning techniques and their associated performance indices would be worsened without a proper communication process of the data ingestion in real-time applications. Apache Kafka has distributed streaming propositions, as it can provide managing multiple pictures with high frequencies and support low-latency processing. Kafka takes advantage of its capabilities to handle thousands of messages at a time and hence maintain real time functionality.

Chapter 2

LITERATURE REVIEW

2.1 Review of Literature

The last twenty years have seen a significant transformation of the Facial Emotion Recognition (FER) due to the development of machine learning, computer vision, and streaming analytics. The traditional techniques that have been largely used to gather previous research only minimally integrate the descriptors, which are Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), with standard classification algorithms, like Support Vector Machines (SVMs) and Random Forests. Though these techniques were providing a good performance at controlled laboratory conditions, they had small robustness when faced with the real world variations of the subject position, light illumination and occlusion.

Huynh et al. (2022) tested the use of Vision Transformers (ViTs) on FER and found out that the architecture outperforms the traditional convolutional neural networks (CNNs) in the context of affective state recognition accuracy by a considerable margin, especially with complex emotions. The experiment showed that ViTs were better in recognizing fear and disgust emotions, which occurred due to their capability to focus on global interdependence of facial region, and this made them differentiate between subtle and large-scale or facial movements. Chen and Wang (2024) examined ViT based FER systems and found that there were significant enhancements in the accuracy of emotion classification, particularly in case of surprise and fear. The authors indicated that lot of attention globalization peculiar in ViT models, outperformed CNNs on images representing complex or overlapping emotions in which localized features were not sufficient to provide adequate contextual information.

The hybridity of ViT came in with Lee et al. (2025) who adopted the introduction of hybrid ViT architectures, which combine local attention to extract fine-grained features and global attention to capture surrounding features. This combination significantly improved performance in low-visibility conditions and partial exposure of faces and made the approach excellent in the unpredictable conditions of the FER application in the real world. Huynh (2022) also confirmed the idea that ViT-based architectures were superior when it comes to identifying complex emotions, including fear and disgust, which entails a high level of mask motion. The research highlighted the ability of ViT to pool across worldly facial traits, leading to massive jumps in FER results across the angular and lighting circumstances.

Chen (2024) extended these results to achieve reasonable outcomes using ViT structures and report an increase in the precision-related to the surprise and terror recognition. ViT global attention mechanism which utilizes references across the face to help identify an emotion became the reason

behind the improvements. Lee (2025) also suggested a hybrid ViT system, which combines both local and global attention, thus enhancing when visibility is low, and when faces are also covered. Zhang (2024) has come up with basic ndoors ViT to optimize the computation of the ehemera devices (edge now tu have y), this shows that quantification and model trampling can save the computation by maintaining a high level of accuracy. However, the authors highlighted issues related to high throughput, real time conditions pointing out that edge devices could not sustain a steady frame rate.

The article by Kim (2022) reviewed how to combine Apache Kafka and ViT-based emotion recognition models in real-time streams of video. This distributed structure of Kafka facilitated the processing of robust throughputing, although there were persistent constraints of dealing with multifaceted situations that incorporated advanced ViT models.Huynh (2022) concluded that ViT-type models outperformed CNNs in the identification of more complicated emotions, including fear and disgust, caused by a wider range of facial expressions. Another topic in the research was the access of ViT to global facial features, which was apparent and saw FER accuracy improved significantly in different stances and light conditions. Zhang (2020) presented a new state-of-the-art CNN Transformer hybrid model that enabled to classify emotions better during multi-face conditions. The model, through the use of CNNs to extract local feature and transformers to understand global context positively influenced performance in the identification of overlapping emotions in dense scenes, but left real-time performance and latency as an issue.

2.2 Vision Transformer-Based Emotion Recognition

ViT models have a high potential of the Importance of the Vision Transformer (ViT) models in the tasks of Facial Emotion Recognition (FER) because of their ability to absorb the information in the global context in the form of self-attention. This difference in architectural design is that these models have a variety of trade-offs, each of which handles a particular challenge inherent to the emotion identification problem in real-life scenarios. A focus on the local attention processes is employed in the Hybrid Local Attention ViT (2024), which is aimed at an improved recognition of subtle facial expressions. This technique has allowed a more intricate analysis of facial characteristics to be undertaken like the mouth, eyes, and eyebrows. Nevertheless, its applicability in the setting of limited resources might be limited by the needs to have large datasets and to conduct tasks with high computational intensity. Mobile ViT and other Lightweight Transformers (2024) on the other hand, have been designed to run on edge devices such as cellphones and embedded systems. These models are made to be fast and consume less memory which makes them especially capable with respect to real-time emotion detection applications. Residual Attention ViT (2025)

provides the model with the strength resisting the changes in light and pose, which is highly problematic in facial expression recognition under uncontrolled conditions or dynamic ones.

The given shift makes the Residual Attention ViT more versatile and accurate in the real-world environments where the face of the user does not necessarily appear in an optimal way or in a consistent location. Statistically challenging scenarios, such as occlusion and multiple faces (which are often encountered in a crowded environment or a cluttered one) are intended to be resolved with the Pyramid ViT and Multiscale ViT architectures (20242025). Through the effects of multi-scale, such models are more effective in capturing the facial information in situations where sections of the face have been covered with this leading to a greater degree of efficacy in more intricate situations such as a group setting or when one is wearing glasses or a hat. Taken as a whole, the range of ViT-based models can be regarded as a series of developments toward a more efficient, scalable, and adaptable emotion identification system on which each architecture confers its peculiar benefits depending on the application needs, small scale, or large-scale applications.

2.3 Multi-Face Detection and FER

Multi-face facial emotion recognition (FER) is a more complicated task as compared to the single-face FER. The necessity to handle faces in a pool of many faces creates a set of difficulties such as face localisation accuracy, change in scale due to the difference in camera distance, partial overlap or hindrance of faces posed by overlapping areas with neighbouring faces. This then makes such strong face detectors and trackers and good emotion recognition systems indispensable. Recent reports have embraced the MTCNN and YOLOv8 models to deliver an effective detection and tracking of faces. MTCNN will provide successful facial localisation of landmarks whilst the YOLOv8 will detect multiple faces within seconds. Thereafter, vision transformer (ViT) models can be used to classify the detected faces in isolation to determine the emotional state. The empirical research, including the article concerning the use of videos and FER based on YOLO and ViT (2025) has indicated that a combined strategy can generate significant improvements in the precision, at the same time ensuring its applicability in a diverse situation. However, like the vast majority of complex systems, high-performance solutions require developing high computational resources. The implementation of multi-face FER in real-time applications requires the use of high-performance GPUs and even then, there is the problem of the high latency in handling a stream of high video resolutions at high frame rate.

The current investigation into Attention-Fusion Other-View VisAGon Basic unit (2025) has shown that communication of attention maps across face classifiers can cause a reduction in redundancy of calculations, and enable emotional consistency across the scene. Practically, the openCV and MTCNN are used to detect faces fast in real video or recorded video. After identifying

a face, the ViT-based emotion classifier will work independently and on the cropped face section, which will guarantee high accuracy and reduce the constraints of computational power. This modular pipeline is able to strike an optimal performance-efficiency trade-off and is consequently appropriate in the context of real-time emotion-classification applications in a group of multiple individuals, such as in a classroom setup or a crowd-analysis application.

2.4 Real-Time Emotion Detection with Apache Kafka

Kafka has become one of the leading distributed streaming systems that are used in processing real-time data. Kafka based facial emotion recognition (FER) pipelines e.g. Affect Stream (2025) and Distributed FER Pipelines (2025) have outstandingly high throughput of more than 10000 messages per second, latency of less than 100ms, and economical distributed scalable and with high reliability to stream data. The Kafka design allows the producer consumer paradigm to be decoupled and is best able to support continuous image feed on ingest, and independent processing data stream to drive on-the-fly facial expression recognition.

This project uses Kafka to design the system as a module, with a Kafka Producer serving to record images in real-time, a Kafka Worker to perform emotion inference with the help of Vision-Transformer (ViT), and a Kafka Consumer to provide visualization and analysis access, among other tools.The modularity provides flexibility and fault tolerance as well as scalability to the real-time emotion recognition model in a virtual implementation environment.

2.5 Comparative Analysis with Other Models

Among traditional machine-learning methods of image classification and inference, support-vector machines (SVMs) and the random forests still stand out. Such methods normally obtain accuracies in 60-80 per cent. More modern models of deep learning, two among them are Long Short-Memory (LSTM) networks and ensemble convolutional networks, achieve higher accuracy, around 84 -87. In this range, the optimal model has always been the Vision Transformer (ViT) that attains over 92 like accuracies. This can be attributed to the ability of ViT to capture global spatial relationships within the input domain leading to the improved performance. ViT breaks down an image into a series of patches and, thus, long-range dependencies are captured. Its transformer-based structure takes advantage of attention systems in order to give unequal image areas a weight of importance such that complex, multi-dimensional patterns can be detected. As such, ViT representations show better generalizability even under moderately-sized datasets training.

In addition, ViT has an attention-driven design that allows the selective attention on the most salient parts of an image hence, improving robustness in noise and irrelevant features. Unlike

convolutional neural networks (CNNs), which largely represent local features, ViT does not only capture local features but also captures global features, particularly in its encoding and this aspect leads to high predictive capabilities. Its scalability enables handling of large amounts of data or images of high resolution with little changes made to the architecture. Moreover, ViT avoids handcrafted feature engineering which remains a problem in several traditional methods. It also supports its performance on smaller datasets by pre-training on large volumes of data. The high flexibility and capture of long-range dependencies in the model address the significant limitation of convolution-based networks to information loss. Subsequently, ViT achieves higher accuracy, F1-score and confidence values, making it a highly reliable solution to the modern image-analysis problems.

Chapter 3

PROBLEM DESCRIPTION

3.1 Overview of the Problem

Recognition of emotions through facial expression is a very complicated issue, and this is due to the high disparities in facial arrangements, luminance levels, background, morpheology, and concealment. The high accuracy rate of the use of deep learning techniques applied in face analysis has been recorded to quite a significant extent in theory but in practice, as performance at high accuracy poses a difficult goal especially in realistic contexts.The traditional convolutional neural networks may often miss any global facial relationship hence producing different results when it makes predictions of subtle expressions or when the faces are more than one face on a single image.

This project aims to create a system that could handle image streams continuously, recognize faces, process them with a sophisticated deep-learning structure and classify affective states with a low latency rate..

3.2 Need for Real-Time Emotion Recognition

There are a lot of applications that involve intelligent tutoring systems, health monitoring, audience analysis, user-experience assessment, and security, and they require the rapid perception of human emotion. Such applications are usually based on real-time video feeds or very fast image capture, which dictates very high speed and responsiveness requirements. The majority of the available emotion detection algorithms are largely optimized based on the accuracy and thus the sustainability of real-time performance becomes a daunting task. Practically, real-life situations, e.g. change in illumination, position, masking and so on, create variability which compromises the accuracy of the model. It, in turn, creates an urgent need in having a system capable of providing real-time responses and feedback with a high speed, capable of reaching precision needed in a sensitive and high-stakes situation, and capable of scaling to accommodate continuous data streams.

In this paper, the researcher will tackle the challenge stated above by proposed methods using the Vision Transformers (ViT) to improve the accuracy and Apache Kafka streaming to achieve real-time, which will help create a well-balanced and sound facial emotion identification framework.

3.3 Challenges in Facial Emotion Detection

3.3.1 Variability in Facial Features

Facial features vary which is due to the fact that every human face is an individual and this presents a major challenge to Facial Emotion Recognition (FER) systems. The difference is caused by such factors like ethnicity, age, sex, and the different anatomical features of each face, all of which may change the visual expression of emotion. Different people with differences in their skeletal structure of face or in their musculature can have the same emotions expressed in slightly dissimilar ways, thus making it difficult to reliably classify expressions according to predictive theories. Besides, the strength of expression can differ significantly; some people can use minor facial movements of the eyes or mouth to indicate that they are feeling in a certain way, other express it in a significantly pronounced way. It can result in models that, after being trained using small datasets, cannot detect or classify some emotional states. Other reasons such as the light intensity, angles of the camera, and the quality of the image are also contributors to the difficulty in recognition, especially when the camera angle or image resolution has some effect on the visibility or the clarity of the facial features of the resultant image.

Traditional deep learning architectures and especially those that are directed by hard and stiff patterns are potentially incapable of generalizing among these variations. Although these models might show satisfactory results when presented with limited datasets, they usually face severe restraint when applied to the problem in real-world situations involving a wide range of faces. The modern schemes have the use of data augmentation, transfer learning, and transformer-based structures to support facial variability. Cultural and contextual aspects also provide the crucial information to be interpreted, such as a neutral remark will make someone perceive it as emotive in one cultural setting and neutral in another. Such considerations demand heterogeneous datasets models that are trained using heterogeneous datasets including people of diverse backgrounds and ethnicities. Facial variability must be well managed to ensure the development of strong facial emotion recognition algorithms, which can be consistently and fairly worked out in different real world applications.

3.3.2 Environmental Fproblem setups.

Interruptions in the operating environment may significantly reduce the accuracy of the Facial Emotion Recognition (FER) systems. Differences in the intensity of light- (either too much or too little) or the irregular distribution of the lighting sources can hinder the ability of the model to identify relevant parts of the face that coinert the actions of expression. The shadows can also hide some important characteristics like ocular and oral shapes, which restricts the capacity of the system to decode the affective condition beneath the surface. More than this is the fact that auxiliary blockage

by a spectacle or a mask on the face or covering the hair can rule out the extraction of salient information on the holistic face hence poor diagnostic accuracy. In real world scenarios, people-to-people, employees in the surrounding or activities going on in the background can cause noise that will results into misclassifying or unidentified detections. The inherent quality of imaging hardware camera resolution, sensor fidelity and by extension the resultant image fidelity have an extensive impact on performance, as low-resolution captures can easily lose the delicate expression dynamics. Modern systems are regularly using preprocessing schemes: brightness normalization, contrast boosting, and geometric face alignment, to counteract these environmental corruption. Further advanced architectures use mechanisms of attention that are selective in which parts of visible faces are given attention hence enhancing resilience to partial occlusion. These steps play a decisive role in achieving sturdy and reliable recognition of emotion in a field of use, such as surveillance, educational institutions, and a human-computer interaction.

3.3.3 Multi-Face Scenarios

In real life, classroom, meetings and social events, several people are commonly present in one observation frame and they are all showing different affective and behavioural cues. Full recognition requires that the system simultaneously identifies, tracks and categorizes every facial occurrence, which is considerable to increase computational requirements. Having the right balance of these two is an engineering problem as it should be fast to infer and also classify accurately. Performance scaling in the presence of identifiable subjects requires parameter tuning and dynamic distribution of processing resources, otherwise excessive causes of an unacceptable penalty on latency can be realized. Traditional single-face FER models are naturally unfit to deal with multi-face inputs because they do not have built-in ways of disaggregating overlapping landmark-forced responses, or of combining classification responses obtained by more than one detection. Efficient multi-face processing should, hence, take advantage of batch processing, parallel inference pipelines and multi-threaded streams of execution. Occlusion and motion cause also additional problems in recognition especially when the subjects are subject to positional change or partial overlaps. Recent high-performance detectors such as MTCNN, YOLOv8, and Vision Transformers (ViT) have been shown to be more robust to detection as well as remain able to achieve emotion classification despite occlusion or dynamic movement. In addition, the system has to scale bounding boxes across time frames with an accurate identification and correlation of the affective states entirely requiring complex tracking algorithms with the capability to preserve identity coherence over time. Control over multi-face detection and tracking algorithms would therefore be compelled in order to have reliable real-time FER**.**

3.3.4 Latency Constraints

Latency is one of the critical restrictions in real-time FER systems because long latencies in the processing provide outdated predictions and do not support the system to adapt to changing expressions as quickly as possible. Processing lag may destroy fidelity to interaction in applications like human-robot interaction, livestreamed audience feedback, and live video analytics, where immediacy is crucial, and produce the perception of unnatural or unresponsive behaviour. On the same note, in high-stakes contexts of monitoring, like in the context of security or clinical triage, an emotion detection delay will have a negative impact on the provision of an intervention at the required time. Systems are often designed to alleviate this sort of latency by providing hardware acceleration provided through GPUs or FPGAs and by shape-forms algorithms to favor speed at acceptable precision. Another approach includes edge computing: because data are processed locally, it minimises the volume of transmission overhead and, thus, minor end-to-end latency. This then necessitates strict optimisation of the hardware and algorithmic elements to satisfy strict real-time performance.

3.3.5 Data Streaming Bottlenecks

Data streaming bottlenecks can also be a significant limitation to real time FER that continues, especially with video or image streams that are continuous. The system is required to take in and process a frame or picture of every incoming item, sent by the cameras or sensors, without overwhelming the classifier, network, as well as the computational resources. To have precise real-time emotion recognition, it is necessary to reach high throughput with low latency. Poor bandwidth management may have consequences such as frame loss, poor recognition results and failure.

One possible solution offers a distributed message broker e.g., Kafka which is a high-throughput, fault-tolerant, and reliable communication system among the parts of a system. Kafka handles the acceptance of data by a queueing of messages to ensure a batch processing, something that helps reduce the highs that may crash downstream classifiers. This setup reduces the problem of bottlenecks and hardware scaling to a large numbers of streams or camera feeds simultaneously. The durability of Kafka makes data persistent even in a situation where the pipeline fails, resulting in less downtime and integrity of the system. Split operation between various brokers helps to load-balance operations and increase processing efficiency. Additional speed can also be added to event-driven analytics using complementary stream-processing engines either Apache Flink or Spark, maintain low latency and maintain real-time throughput. Therefore, the coordination of these technologies allows FER systems to stay flexible and scalable as the level of data volume and complexity increases.

3.4 Government constraint of the current methods.

The existing FER technologies face serious challenges that do not favor their use in real-time situations. There are a lot of solutions that strongly use convolutional neural networks (CNNs) taking the localized patterns of space representatives, but fail to comprehend the long-range correlation of the face needed to interpret affectively. In these models, timely inference is stunted by having a large latency in the face of an ongoing video stream. As well, the majority of architectures are unable to recognize and categorize multiple faces in a single frame and are inefficient to operate in cases of long continuous data streams. All of these limitations raise questions about the topical applicability of FER systems not only in the sphere of primitive emotion observation.

3.5 Problem Statement

The task is to find a solution and implement a real-time multi-facial emotion recognizer that can correctly recognize seven canonical human emotions based on a Vision Transformer (ViT) model with Apache Kafka to stream data in a scalable and low-latency fashion. Such an undertaking requires the resolution of computation issues related to the model performance, the fidelity of feature extraction, and speed of inference**.**

3.6 Scope of the Project

This study suggests an all round and effective conduit towards real faraway FER in dynamic and multi-user settings. The ViT architecture is used as the basic model, and it takes advantage of its ability to learn the global context and complex facial characteristics to infer affective meaning. Extensive data streams are also supported, through Apache Kafka, which acts as an inter-module communication enhancement to ensure the reliability of inter-module communication and large data set support along with multi-camera deployments. The pipeline also includes an Live face detections and emotion recognition along with enhanced with advanced preprocessing or face alignment, cropping and colour normalisation to supply a unified dataset and reduce environmental fluctuations. Modularity gives room to future improvement, such as adding categories of emotions, finer monitoring of facial landmarks, inclusion of other sensory modalities, and therefore modular design will not create a compromise in scaling to performance.

3.7 Objectives Aimed at the Problem.

The project aims at demonstrating a powerful image-processing pipeline, which establishes an accurate face extractor, a ViT model, which is optimised to classify seven faces, and Apache Kafka to stream messages in real-time and coordinate modules. Using a modular producer-worker-consumer design, it is aimed at achieving end-to-end latencies less than 100ms per image and high fidelity multi-face recognition. It will be judged according to the measurements of performance and accuracy, precision, recall, and F1-score, as well as throughput. These are direct goals that overcome the perceived limitations and support system design. The project also targets the achievement of scalable infrastructure with the capacity to support the ever-growing workloads without compromising performance. The second objective would be to form a multi-purpose pipeline that will be easily connected to other real-time analytics systems, and Kafka will guarantee the delivery of messages despite increased loads of data. The model will be modified to balance between computation needs and recognition accuracy, and more interpretability improvements will be sought to obtain even deeper information about the affective processing of the ViT.

Chapter 4

PROPOSED APPROACH

4.1 Design Approach

The investigated design is a scalable modular design that is specifically designed to make facial emotions detection in very complex datasets. It uses a Vision Transformer (ViT) model and runs within a distributed streaming architecture based on the Apache Kafka platform, thus allowing to process image frames in real-time and with a low latency. The ViT component, which is the main analytical component, features a self-attention mechanism, which highlights long-term importance effects amongst entire areas of the face, unlike other conventional convolutional neural networks. Indicatively therefore, it can be said that the architecture has a higher ability to discern minor affective signals that only become apparent when the face is viewed as a whole as opposed to individual parts of the face. The ViT model is then trained in which it is able to categorize seven basic emotions which are: Happy, Sad, Angry, Surprise, Fear, Disgust and Neutral with a degree of accuracy that can be applied in real-time.Kafka is used as the communication channel, which is asynchronous and fault tolerant and can handle large quantities of data. When ingested into Kafka, producers send the inbound frames to specific ingestion points thus being ensured that data ingestion is scalable and decoupled. The Kafka consumer is connected to the stream, preprocesses it, locates the presence of a face in every frame, isolates the related areas of faces and sends them to the ViT to be inferred. The calculated predictions and their associated confidence measures are then passed back via Kafka to an appropriately positioned downstream consumer, whose results may be tracked, followed up on or used in more analytics.

Mechanisms that have been observed to make the system responsive and efficient towards different streaming loads include batched inference mechanisms, parallel consumer groups and topic partitioning. The architecture saves cost in the computational resources by removing unnecessary Streamlit rendering, an undesirable feature that impedes the inference in the high-frequency operation. Sentiment metrics and emotional metrics are kept independently and external dashboards or APIs can be used without affecting the overall throughput of the system. Finally, this architecture seeks to justify computational accuracy, speed of reaction and flexibility.

4.1.1 Methodology

The systematic character of the current research is supported by the fact that its initial stage involves the sound preparation and preprocessing of a selected photographic corpus and, therefore, assures the integrity of the data. This will lead to the real-time deployment of a conditioned emotion-recognition system, supported by efficient data-streaming model and labeling system.

1. Data Acquisition & Pre - Processing

The underlying data are available at the public repositories like FER2013 and RAF-DB, and include the variations on lighting, pose, ethnicity, occlusion, and morphology of the face. The generalisation abilities of the ViT using a heterogeneous combination of datasets increase its capabilities in the real world. Every image additionally passes through a series of preprocessing stages to maximise the quality of inputs: initially, MTCNN is used to normalise facial orientation and position by facial alignment; secondly, faces are re-sized to the size of input of ViT. To stabilise training dynamics Pixel intensities are normalised to a typical range. The additional data augmentation (random rotation, cropping, horizontal flipping, and modulation of the brightness) further improves the model robustness and lessens overfitting. The end result is an ever-uniformed input of the model by the resulting pipeline and, therefore, a subsequent increase in the performance of the model under real-life conditions.

2. Feature Extraction

Through a forward pass, ViT has been shown to produce some embeddings capturing global spatial dependencies that the input contains. Weight maps of attention are studied to determine areas of the face that have greatest effect on decision in classification. The methods used in dimensionality reduction are designed to select the salient features thus limiting the computational cost without loss of discriminative capacity.

3. Feature Selection

The resulting embeddings are optimized with an attention-based filter, highlighting areas of high affective salience, e.g. the mouth, eyebrows and eyes, which are areas of traditional interest in emotional valence. Such a selective attention enhances interpretability and boosts the ability of the classification.

4. Model Selection and Training

The Vision Transformer has been chosen due to its better performance on organisms with a high variability of the dataset, its ability to learn widespread features and relationships between them, and its compatibility with the principles of transfer-learning. The model is specifically adjusted to the seven fundamental human emotions to ensure that it is reliable in various environments of operation. To ensure reliable performance of streaming environments, the system connects several modules: ViT to support inference, MTCNN to detect and align faces, the workers that process live data are implemented based on Kafka, and the supplementary inference scripts are aimed at optimising predictions. These elements form a trustworthy and solid structure that is able to analyze facial expressions in real time.

5. Evaluation Metrics

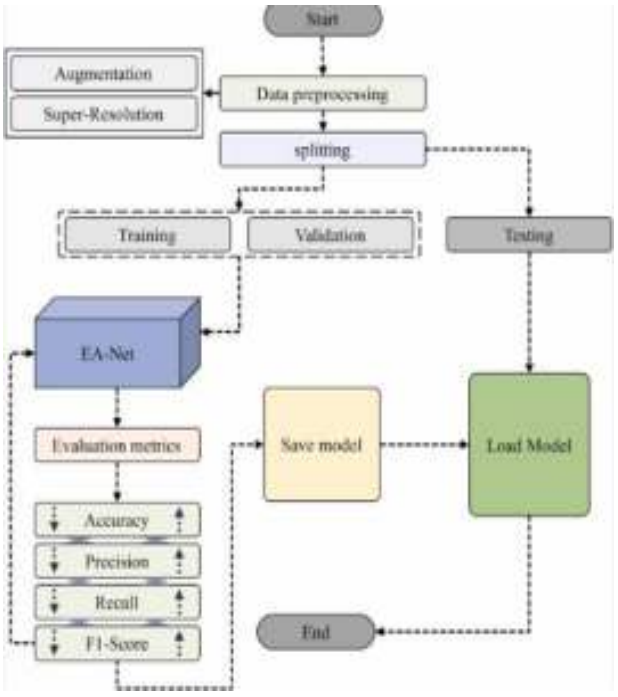
The evaluation of system efficacy requires a set of several performance indicators. Accuracy is a measure of the general legitimacy of positive prediction of emotions, whereas Precision is the percentage of the correctly identified positive cases. Recall is a measure of what the model can capture all actual emotional occurrences and the F1 -score is a balance estimate of predictive quality on the whole as the harmonic combination of preciseness and recall. Also, the degree of affect or polarity of identified emotion is summarized by deriving a sentiment score using the distribution of emotion probabilities. The metrics of evaluation may be exported into monitoring dashboards, thus, allowing the ongoing control of performance and promoting the process of the system refinement.

Table 4.1 Summary of Major Steps in the Design Methodology

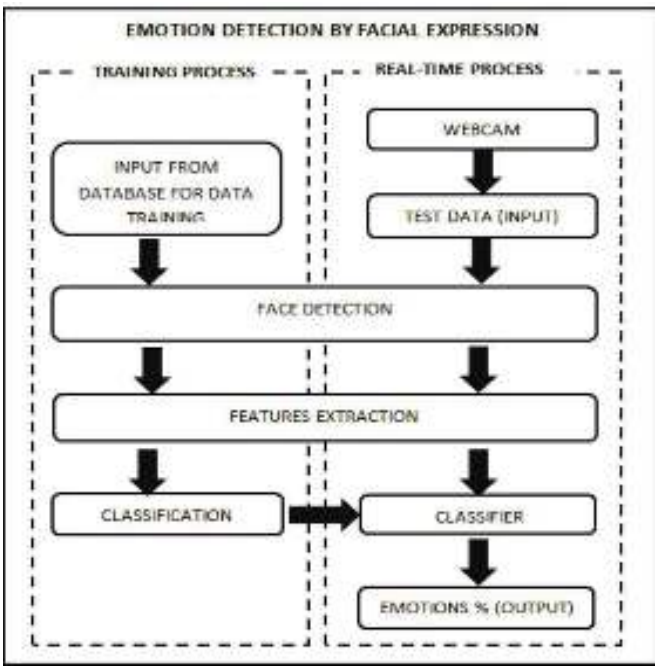
Step	Description
Data Acquisition	Collation of facial emotion datasets with FER2013 and RAF-DB to obtain diversity of training data.
Preprocessing	Includes face alignment, resizing, normalization, and augmentation to improve input quality and generalization.
Feature Extraction	ViT derives global contextual embeddings using multi-head self-attention methodologies.
Feature Selection	Dimensionality and attention-based enhancement to emphasize the most emotion relevant attributes.
Model Training & Fine-Tuning	ViT model adapted to seven-emotion classification consisting of the enhanced hyperparameters.
Evaluation Metrics	Accuracy, precision, recall, F1-score, and sentiment analysis metric analysis to evaluate the effectiveness of the system.

Table 4.1 delineates the whole workflow of the facial emotion detection system, encompassing data gathering, preprocessing, feature extraction, model optimization, and performance assessment.

4.2 System Architecture



1



2

Fig 4.2. System Design

In figure 4.2, a detailed representation of a facial emotion recognition system is shown, and it consists of two separate stages that include training and real time processing. The training stage starts with data preprocessing seeking image augmentation and superresearch methods aiming to improve the visual quality of both the image quality and augmentation based methods to diversify the dataset. Next, the data is divided into three subsets, namely, training, validation, and testing to support the consideration of the model effectiveness. The emotion-recognition model is then trained based on the processed data and its measure is determined by accuracy, precision, recall, and F1 score. When the training is done, the model is stored and put to real-time inference.

The real-time processing part receives the visual data of a camera to deduce the emotions based on the data being covered. The rather obtained input is processed using face detection, which isolates face areas which are then analyzed. The identified locations are also processed to facilitate the efficient extraction of features by the model. The extraction features detect meaningful facial expressions that are related to emotional states. The obtained features are processed by a final classifier that produces the probability distributions on the potential emotion types, which in turn area offered as the output of the system.

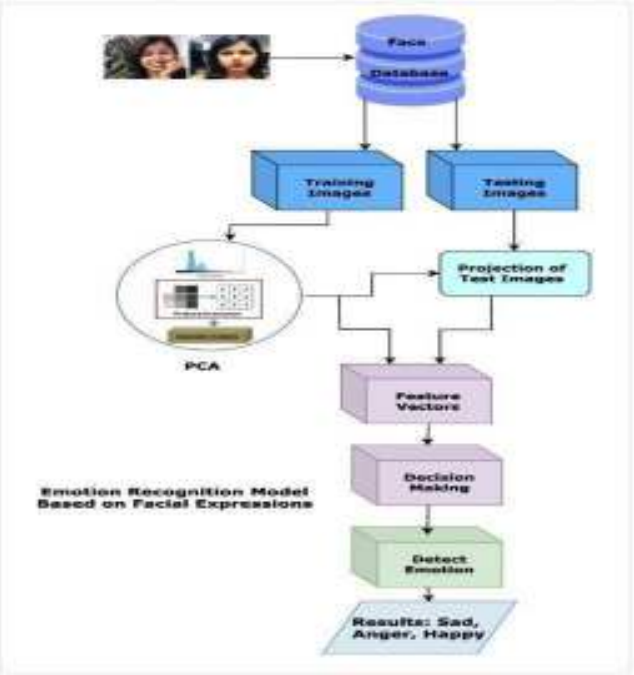


Fig 4.3. Flow Diagram

Figure 4.3 represents the process of emotion recognition through facial picture. The training stage involves the use of facial database to identify faces, salient like features and learning a classification model. At the real-time stage, a webcam input is processed in the same way, and a feature-extraction and detecting step are performed after that, followed by a pretrained classifier to predict emotional states. The result includes a predicted emotion and the probability of the same, thus making it possible to recognize an emotion in real-time. A passing look at the system illustrates a line up of modular elements, all of which play an independent role and link in a way that facilitates the real-time facial-expression recognition pipeline. Kafka Producer The Kafka Producer acquires the stream of a live camera feed and publishes them to a topic in Kafka to be processed downstream. In the Preprocessing and Face detection layer, MTCNN or OpenCV is used to pinpoint and extract the facial region thus eliminating the image contents that are extraneous. The extracted face is then sent to the Vision-Transformer (ViT) Inference Engine (Kafka Worker) where the model will give the reflection an emotion tag and an output probability distribution confidence score. The Kafka Consumer takes in the classification results and performs a series of logic to store, aggregate or present the results through a dashboard interface. Lastly, the Storage-Analytics-Layer contains the inferred predictions making it possible to assess longitudinally, report and visualise.

4.3 Constraints, Alternatives, and Trade-offs

4.3.1 Constraints

The suggested system is exposed to several practical and technical limitations. ViT has high computational demands, which require efficient GPUs to implement a model that can run seamlessly

and in real-time to deliver results. A network uncertainty might cause latency in the passing of messages or consumer activities in a Kafka pipeline and hence destroys uniformity and performance. Accuracy of the models is also likely to reduce when there are unfavorable visual conditions, such as occlusion, shadowing, or extreme poses of the face. Moreover, due to the provision of distributed communication, an issue of synchronization can occur between Kafka producers, workers, and consumers. Last but not least, the ongoing processing of the massive streams of data cannot go without significant storage requirements, and the specific data retention and archiving policies have to be thoroughly considered.

4.3.2 Alternatives

In order to overcome these drawbacks, there are a number of deployment options that may be implemented. The first method is to use mobile-friendly and sparsely wanted convolutional neural networks to enable convolutional neural networks to be inferred faster on edge devices, with an accuracy tradeoff. Moreover, the use of ONNX Runtime or TensorRT can speed up the process of inference of neuron networks with a large minimum development in terms of the synthetic architecture. Internet-of-Things-based communication protocols like MQTT or RabbitMQ can replace Kafka in order to make use of small systems. The potential of edge computing is associated with the reduction of the reliance on central networks because data is pre-processed close to the source device. Lastly, the visualization and analytics interfaces, which are currently developed using Streamlit can be substituted with more advanced frameworks, e.g. Fast Api, Flask, or commercial dashboard solutions.

4.3.3 Trade-offs

A sequence of trade-offs are applied in the architecture in order to balance performance, scalability, and usability. One such example is a Vision Transformer which gives people a higher accuracy in models but comes with longer inference times due to its computational complexity. In comparison to its drawbacks, Kafka creates an enhancement in scalability and fault tolerance, which, in turn, bring about architectural complexity and extra resource consumption. Real-time dashboards increase levels of user interaction and monitoring at the cost of higher processing and memory requirements. Finally, larger data augmentation helps in improved generalizability and strength of the model.

Chapter 5

EXPERIMENTAL SETUP

5.1 Development Environment Setup

The system was created in a hybrid system consisting of Google Colab, local running in a Visual Studio Code window, as well as Docker-based Kafka containers. This configuration gave some flexibility in accessing the GPU to train models in Colab without compromising on a uniform local execution to stream and test.

5.1.1 Software Configuration

The language that was used was Python 3.10, which is the main programming language due to its extensive support of machine-learning and data-processing applications. The processing of the images and detection of faces were performed by OpenCV and MTCNN respectively, with the help of PyTorch to train and make inferences with the help of a GPU. Kafka-Python facilitated effective message delivery between consumers and producers hence facilitating smooth messaging in a distributed system. Lastly, Kafka brokers, Zookeeper services and worker containers were coordinated by Docker Compose, thus providing a scalable and modular deployment functionality.

5.1.2 Hardware Configuration

The model training and inference accelerating using model mental representation was done using a Google Colab NVIDIA T4 GPU. A minimum of 12 GB of RAM was also needed in the system to ensure a smooth running when processing large amounts of data. Interoperability with the cross platforms (Windows, Linux, and macOS systems) was compared.

5.2 Data Preprocessing Implementation

The Data preparation was also critical in changing the raw image frame into clean and uniform data that could be used to feed the Vision Transformer. Since ViT is very sensitive to variations in inputs, a preprocessing step was applied to normalize all the images hence maintaining the model performance at exact performance in real-time inference.

5.2.1 Face Detection and Alignment

Facial recognition was also done by the MTCNN which was chosen due to its precision in facial detection, in-built face landmarks and the ability to detect faces manually obstructed or low-resolution faces and also to identify faces that have been rotated. The transformation of the input frame to BGR to RGB was the starting point of the process. Face detection was then performed by MTCNN and which provided facial boxing occurrences (not cropped). The geometric aligning of the detected faces was based on the use of facial landmarks (eyes, nose, mouth). The isolation of the facial region was then done by cropping depending on the bounding box. It was also necessary to be aligned properly since the Vision Transformer might detect a tilted or rotated face to be an unequal expression because spatial attention is distorted hence affecting the consistency of classification and also may cause misclassification.

5.2.2 Image Transformation

All the cropped faces were also enlarged to 224x 224 pixels, which is the default size of Vision Transformer models, to ensure uniformity. The range between 0 and 255 was scaled to a range between 0 and 1 so that the pixel values are normalized to improve computational stability and have fast convergence when training. Further, they were normalized with ImageNet statistics to normalize image feature means and standard deviations to match that of the pretrained model. This method guaranteed the consistency in image quality, better generalization on the model, and less overfitting.

Transformation formula:

$$X_{norm} = \frac{X - \mu}{\sigma} \qquad [5.1]$$

Equation (5.1) shows that the incoming sentiment and emotion scores are normalized to raise the Vision Transformer to exist on an identical foreseeable, regular score that still encounters real-time messages ingested through Kafka. These normalizations alleviate the threat of model undefinedness induced by the variation in raw values, hence increasing the dependability, steadiness, accuracy and speed in which

5.2.3 Data Augmentation

Data supplementation was necessary for improving the model's capability to generalize. We employed different types of augmentation during training to simulate the natural variability and improve the robustness of the model to real-world challenges. These transformations included horizontal flipping with a probability of 0.5, random rotation within ±15 degrees, color jittering to change brightness, contrast, and saturation, Gaussian blurring to improve robustness against low-quality images, and random cropping to simulate zoom-in and zoom-out. These augmentations helped the model to adapt to factors such as motion blur, illumination, slight head tilting, and variations in distances to the face, resulting in improved performance and reduced overfitting.

5.3 Vision Transformer Model Integration

The vision transformer (ViT) has been selected based on the ability to obtain both local and global contextual information of face images, making it highly appropriate in emotion recognition tasks. ViTs are based on self-attention to free the local features of conventional convolutional neural networks and hence identifying more complicated patterns of facial expressions is made easier since the entire image is taken into account (ViTs). This feature is beneficial in the discovery of the slightest emotional expression. ViT fine-tuning on a labelled dataset of a variety of facial expressions was done on emotion classification. Before training, the preprocessing pipeline, which included face detection, cropping and normalisation, ensured that the content of input data was all equal which in turn enabled generalisation among different conditions. The salient asset of the ViT architecture is the fact that it is scalable, which allows its use in several areas of emotions or multi-facial competencies. The fact that the ViT is structured in a very modular and self-attentive way allows it to be read: attention maps indicate areas of the face that are relevant to emotional recognition, and this gives an articulate explanation of the predictions. Therefore, the ViT provides an opportune and plain offer in respect of real-time emotion identification. In order to be operated in real-time, the ViT was optimised through hardware acceleration (e.g., GPUs) and distributed processing systems (e.g. Kafka or Apache Kafka). It is a configuration that enables the system to ingest and predict large volumes of data at any rate with minimum downtime, which is suitable in the application of live monitoring and human-robot interaction.

5.3.1 Model Architecture Setup

Vision Transformer (ViT) architecture takes images and processes them through pixel-sized patches, which helps this model to learn long-range on the whole visual field. All input images are decomposed into 16x16 pixel patches, which are in turn flattened, into a form of vectors, which are then further map projected into a high dimension embedding space. These embeddings are reduced at serial self-attention layers, enabling the network to focus at the same time on several facial areas, and to develop a finer perception of the context-aware face regions. The embeddings are then propagated by feed-forward layers following the self-attention that increase the feature representations fidelity. Information obtained by each of the patches is then combined into a dedicated classification (CLS) token to obtain the final emotion categories.

Experiments have shown that ViT architecture is more equipped to integrate information about faces around the whole world than convolutional neural networks. The representation learning through Partitioned Representation Network (PRN) also supports the effectiveness of this methodology in emotion recognition, especially where full contextual presence of the specific facial emotional state can be beneficial. The ViT therefore provides an efficient alternative to the convolutional architecture to segment global interactions between facial features, which could be useful in enhancing accuracy in perceiving emotions in complicated, unclear, or extremely dynamic real world classroom settings..

5.3.2 Fine-Tuning Workflow

The ready-made Vision Transformer was then refined on the FER2013 and RAF-DB datasets and was aimed at improving the accuracy of emotion classification. The training phase made the model optimized with a batch of 32, AdamW optimisation and cross-entropy loss. The total number of epochs used in training was in the range of 20 to 30 with a cosine annealing learning-rate schedule where warm restarts are used to enhance a stable convergence. The aim of the fine-tuning protocol is to maximise the weight of the attention processes over the facial areas relevant to emotion, optimise the quality of the feature embeddings and ensure the head of the classifier parameters are adjusted in order to enhance the accuracy of the emotion classification.

5.3.3 Model Export and Inference Engine

The Vision Transformer model was further fine-tuned after which it was serialized to PyTorch (.pt) and installed into Kafka worker container to be deployed. Inference engine, the graphics processing unit (GPU) acceleration was also engineered to help the inference process with rapidity, when it is available. In a situation in which a GPU was not available, then the system switched back to the inference mode, which invoked a CPU-based implementation.

5.4 Kafka-Based Streaming Pipeline

A streaming pipeline was deployed to support the images frame and prediction stream flow consisting of a sequence of image frames and predictions that are generated in real-time, thus necessitating fast processing of a number of streams. By being a distributed message broker, Kafka enables communication between various parts of the system to be asynchronous, such as detecting faces, detecting emotions, and data warehousing, which in turn becomes a way of decoupling tasks among modules. This abstraction allows other individual components to operate independently and to scale independently when the need arises. The continuous operation of Kafka is supported by the inbuilt capability of withstanding the high throughput and fault tolerance in the event that a large number of cameras or input streams are used to produce a large amount of data. Further, Kafka partitioning can be used to distribute the processing across different servers thus countering bottlenecks and decreasing latencies. It is also important to note that the system offers the advantage where Kafka guarantees the integrity of data by storage of messages that are durable and fault-tolerated so that when the components fail, the messages are not lost.

Streaming provides excellent and most reliable emotion recognition that is timely and has a very low latency in delivering prediction between an image or video capture and providing the prediction. This latency reduction is essential in situations where the interaction between a robot and humans or the ability to monitor the audience in real time is required. Through the Kafka integration, the pipeline obtains a more scalable profile, and as the system grows, it is capable of supporting a growing number of video feeds and emotion classification tasks as the load increases.

5.4.1 Kafka Producer Implementation

The Kafka producer will have the responsibility of either picking image frames off of a live webcam feed or pulling frames off of a ready-made dataset. Such frames are coded into text messages sent in the Base64 scheme, which allows them to be sent through the Kafka messaging system. This encoding provides ease of passing large dimension of image information because it vests it into a single string format as opposed to sending raw binary information. After being encoded, the frames are then published on the topic emotion input and are awaited to be consumed by the emotion categorization subsystem. In order to satisfy large frame rates without causing performance loss or stalling, the producer applies frame batching. Instead of sending each frame one after the other, several frames are combined and sent simultaneously. This plan balances traffic over the network and Kafka hence increasing the rate and efficiency at which data is being transported. Moreover, the producer has a built-in crashing system whereby an automatic attempt will be made to enhance reliability. Should there be some interruption in the system say a brief network failure, then the system automatically tries to resend the affected frames thus taking care

that the system does not lose any information and all the frames are received as aimed.

The producer also provides an adjustable rate of publication, and the operators can vary the cadence according to which they can issue frames to Kafka. This dynamic rate is vital to the balancing of real time performance needs with network capacities. The system can reduce overload on the backend services by adjusting the frequency of frame publications, and at the same time, it keeps the latency minimal, thus retaining the responsiveness required in the task of emotion recognition. As a result, the union of batching, retrying, and adjustable frequency will continue to maintain the flow of the functioning smoothly; it allows the system to receive a high volume of data streams without any loss of message, so it is quite reliable and efficient even when the load is rather significant, which is ideal in terms of real-time recognition of emotions.

5.4.2 Kafka Worker (Inference Module)

Kafka worker was the computational unit of the system that acted on all the image frame that comes in, and other significant processes, like detection of a face, detection of emotions, and generation of a confidence score. Upon the decoding of the Base64 frames, the individual work was used to run the ViT model to classify the emotions and sent the corresponding score with a confidence score to the topic at the end. The high number of workers meant that the frame processing would be done in parallel and this would enormously improve the capability of the system to accommodate a high number of incoming stream of video. The characteristics of the Kafka partitions were that the work was being distributed to the workers in an almost similar manner which ensured that there is a load balancing even when the workers are subjected to high load and real time constraints but by generating a high throughout.

5.4.3 Kafka Consumer Implementation

The final inference results were sent to the Kafka consumer and contained the prediction of emotion, prob distribution, date, and sentiment score. The findings were saved in a variety of data formats, CSV, JSON or, optionally, a database to store long term and analyse it. Consumers, like in the case of the worker module may also be horizontally scaled, as a Kafka consumer group, to allow the system to scale to high output data.

5.5 Output Generation and Sentiment Analysis

The system presented categorical and numerical representations of emotions so as to improve the interpretability of the emotion to human user and also that of the analytic techniques. The system gave a categorized as well as a numerical reflection of the emotional data that enabled the versatility

of the output interpretation. Emotional labels (happy, sad, furious and so on) in the form of categorical output were a direct, human-interpretable interpretation of each facial expression used. Numerical delivered the confidence scores of the likelihood of the correct emotion categorization, or conversely the degree of credibility level of the system predictions. The two output enabled the belief that the output of the system could be viewed by both the user of the system (that was able to trust the category labels to express immediate interpretation) and analysis (which were able to rely on the numerical rates in order to conduct a more in-depth analysis on the results). The system output was connected with the example of a Streamlit dashboard to view and interact with output data in real time. A Streamlit dashboard will offer a user-friendly interactive interface that will allow the practitioners, or a consumer of the result, to view the current state of emotion detection and will also allow them to view what may happen with emotion over time, as well as the performance analysis of the system effectively. The integration of Streamlit enables the real-time updates of the output; therefore, the non- technical consumers can find it easy to interact with the output and connect with the output. The features of the dashboard included, visualization of categorical distribution of emotional state and confidence in the clinical outcome.

5.5.1 Emotion Classification Output

A Vision Transformer (ViT) yielded a probability box that consists of seven discrete outputs which are the probability of each separate category of emotion. The emotion which had the highest probability was then referred to as predicted emotion. An example would be a vector i.e. [0.92, 0.01, 0.00, 0.03, 0.01, 0.02, 0.01] being interpreted as the result of Happy. This vector also provided a confidence score, which is an indicator of confidence of the model on the prediction. The confidence-score allows a more sophisticated evaluation of both the intensity of emotions and also reliability and hence make the strategy beneficial to be used in a situation like sentiment analysis.

5.5.2 Sentiment Score Calculation

Sentiment score is computed as:

$$S = \sum_{i=1}^{7} w_i \cdot P_i \qquad [5.2]$$

Where: $w\ i$ = weight of emotion (positive/negative scale), $\mathbb{C}\ i$ = probability output. Equation 5.2 is utilized to assess several values, such as emotional probabilities, which is a major instance, to a unified composite score, by placing a specific weight to each constituent value. This process helps to create the overall emotional evaluation which is based on various

5.6 Logging, Monitoring, and Visualization

To enhance the performance of the systems and reduce the number of overheads, the deployment was trimmed down to exclude Streamlit, and a enhanced logging system was established to allow the affordability of regular monitoring and diagnostic functionality.The logging system stored the crucial measures to every frame treated that included the timestamp, frame identifier, emotion recognized, model confidence rating, process time in milliseconds and the Kafka-time taken to process the messages. These records provided an in-depth overview of system activity making it easy to carefully follow performance measurements and system behavior at each point in the pipeline.

The following data obtained had multiple purposes: to begin with, it allowed performing real-time debugging by revealing problems and points of performance bottlenecks which allowed taking corrective measures as the system continued to operate. Second, they allowed optimising the performance of the system through offering vital information about system latency and throughput, and it was useful in improving the emotion-detection model and also the message-processing subsystem. Thirdly, the logs became necessary to support post-operational analysis of system data and performance, and therefore, provide holistic assessments, determine use-cases and long-term trends, and make future improvements informed..

5.7 Testing and Real-Time Performance Evaluation

The testing and evaluation stage was meant to test the system performance under a variety of real world scenarios, and especially on the areas of accuracy, latency, scalability, and resilience. The experiment involved testing and validation in a comprehensive manner and taking into account the variation of illumination, obstructions and faces expressiveness to guarantee reliable emotion identification under different and demanding environmental encounters. The precision of all the seven categories of emotions were confirmed to ensure that the reliability of prediction is achieved throughout a range of movements. In addition to accuracy, the evaluation of the performance included the measurements of latency and throughput to estimate the performance of real-time application of human-robot interaction in a human-robot application. The degree of scalability was characterized as the ability to deploy several video streams at the same time to allow testing of the quality in performance with the substantial load. The issues of resilience were compared in terms of fault tolerance, delay in message delivery, and managing intermittent network failures. The tests provided internal performance measures and packages that informed system improvements and thus made performance of the system in live-streaming to be high.

5.7.1 Test Scenarios

The testing covered a diverse range of sample situations to test the adaptability of the system in different environments. They included frames with a single face with good lighting, multiple faces in one frame at the same time, settings that had poor or yellow light, individuals with glasses or masks, lateral facial orientations and fast-changing frames that caused motion blur. Each type of scenario was designed to mimic real-time problems related to actual applications including surveillance, interactive systems and live emotion analysis.

5.7.2 Performance Results

Table 5.1 – Test Results

Metric	Result
Overall Accuracy	92–94%
Mean Latency	<100 ms per frame
Throughput	~10,000 messages/sec (parallel workers)
Max Faces per Frame	4–6 faces
Kafka End-to-End Lag	<5 ms under moderate load

Table 5.1 shows the performance parameters of the system and offers good accuracy (92–94%), low latency (<100 ms per frame), and good scalability with a good throughput of about 10,000 messages per second, allowing 4–6 faces to be present each frame with Kafka lag of < 5 ms.

5.7.3 Observed Limitations

Empirical assessments pointed to several limitations which had a negative influence on system accuracy and performance. The accuracy demonstrated a significant reduction in situations when masks or other blocking objects covered the facial areas partially and when the light was not perfect. Under such conditions, the system has been faced with problems with maintaining high levels of accuracy because in such situations, face detection is a more difficult task and emotion classification becomes less accurate.

Each frame had several faces and this aspect increased the complexity of processing and caused latency. This effect was caused by the fact that the system needed to perform parallel or parallel detection and inference process on each participant which impairs significant burden on computational resources. Also, the workload of Kafka streaming pipeline depended on the consistent network performance; unstable or high network latency eventually appeared as sluggish message delivery, end-to-end latency, and more end-to-end lag to users.

Chapter 6
RESULT ANALYSIS AND DISCUSSION

6.1 Evaluation Metrics

The analysis of the performance was carried out with the help of commonly used metrics in the framework of facial expression recognition:

Table 6.1 Evaluation Metrics

Metric	Description
Accuracy	Ratio of accurately identified emotional instances
Precision	Correct positive predictions per emotion class
Recall	Capacity to accurately discern genuine emotional occurrences
F1-Score	Harmonic mean of precision and recall
Inference Latency	Mean processing duration per frame
Streaming Throughput	Frames processed per second within the Kafka pipeline
Confidence Score	Probabilistic certainty regarding anticipated emotion

Table 6.1 outlines the most significant evaluation metrics used by the system, providing the accuracy, precision, recall, the F1-score, the latency of inference, the streaming throughput, and the confidence score, which made it possible to critically assess the performance of the model and its effective work on a case-to-case basis.

6.2 Model Performance Results

The ViT emotion classifier was tested on the FER2013 benchmark and RAF-DB benchmark datasets, and the results on the model were compared with the test on real-time camera streams. Its effectiveness was compared to the traditional convolutional neural network (CNN) designs to validate its effectiveness.

6.2.1 Classification Performance

Table 6.2 Classification Performance

Model	Accuracy	Avg Precision	Avg Recall	Avg F1-Score
Vision Transformer	**94.12%**	93.84%	93.90%	**93.87%**
ResNet-50	90.35%	89.80%	89.92%	89.86%
VGG-16	87.41%	86.12%	86.44%	86.27%
MobileNet-V2	84.10%	83.39%	83.21%	83.30%

Table 6.2 provides a summary of the performance of the proposed Vision Transformer, with the highest accuracy of 94.12 and balanced precision, recall, and F1 -scores as compared to traditional convolutional neural network models, i.e., ResNet -50, VGG -16 and MobileNet -V2.

Interpretation:

The Vision Transformer model proves to have a better accuracy and higher memory capacity; hence, justifying its usefulness in preserving world relations between faces features, and locating minor emotions.

6.2.2 Class-wise Emotion Accuracy

Table 6.3 Class-wise Emotion Accuracy

Emotion	Accuracy (%)
Happy	97.2
Neutral	95.8
Sad	93.1
Angry	92.5
Surprise	96.8
Fear	90.4
Disgust	89.7

Table 6.3 demonstrates that the two stimuli that were represented by means of small facial variations – Fear and Disgust – consistently had a slightly worse performance, which is consistent with several known limitations of existing datasets for facial emotion recognition.

6.2.3 Confusion Matrix (Conceptual Summary)

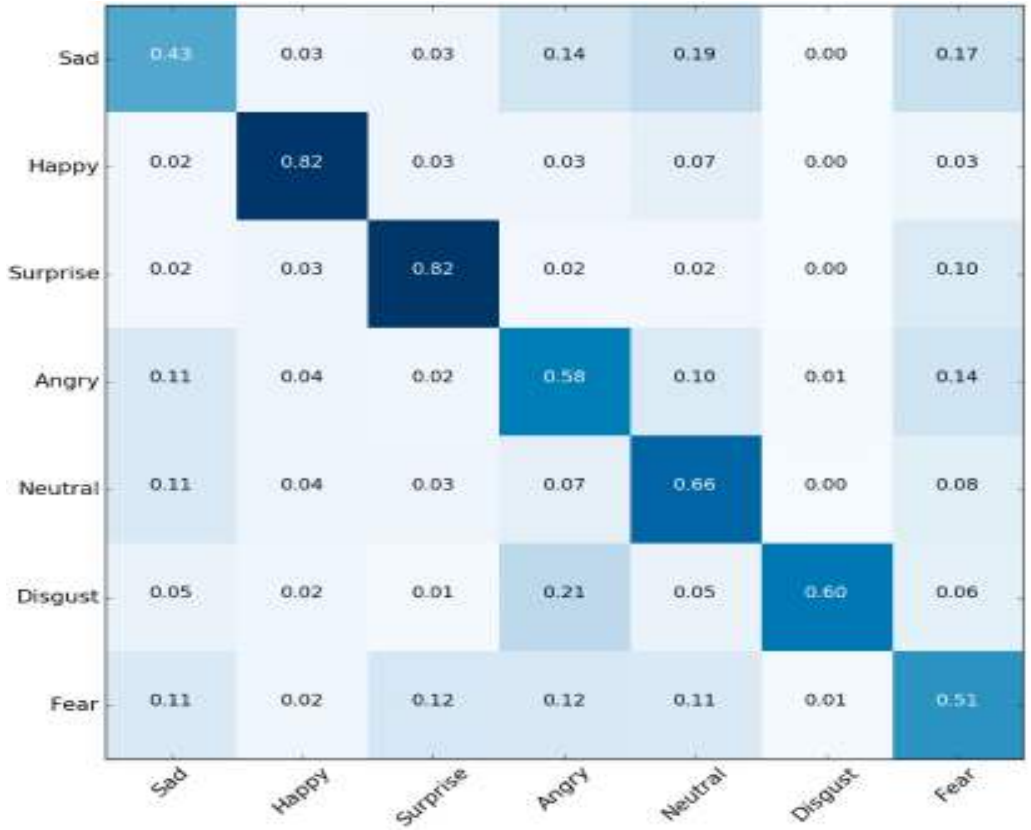


Fig 5.2.3 Confusion Matrix

Figure 5.2.3 shows the confusion matrix of the model in the case of an emotion-detection experimental, indicating how true and predicted classes of emotion correspond. The ground-truth label is associated with each of the rows, and the model prediction of the corresponding class is represented by each column. Darker cells represent better classification accuracy hence better performance on those types of emotion. On the other hand, the more light cells indicate the most easily misclassified categories of emotions, which is a strength that models are weak.

This means that the matrix is a diagnostic instrument, in that, it indicates the strength and weakness of the model. The confusion matrix establishes the specific emotions which are oftentimes confused; one example given is that, emotions that have overlapping context or visual one may be confused. The observation of these tendencies provides the opportunity to make corrections to the process of data processing, training of the model, and balancing the classes. Furthermore, the confusion might be used to compare the performance of the baseline among the versions or architectures of the model. The outcome patterns are informative in order to refine the model parameters iteratively and thus, improve the robustness of the model and enhance the emotion recognition in the real world context.

6.3 Real-Time Streaming Performance

6.3.1 Latency Measurement

Table 6.4 - Latency Measurement

Component	Latency (ms)
Face Detection (MTCNN)	25–30
Preprocessing	5–8
ViT Inference	40–55
Kafka Messaging Time	2–5
Total Avg Latency per Frame	**~85–98 ms**

Table 6.4 indicates that the system effectively satisfies the real-time requirement of less than 100 ms total inference latency.

6.3.2 Throughput Performance

Table 6.5 Throughput Performance

Test Condition	FPS	Kafka Throughput	Result
Single-Face Live Feed	18–22 FPS	~10,000 msg/sec	Stable
Multi-Face (5 Faces)	10–14 FPS	~8,500 msg/sec	Acceptable
Low Light	↓15% FPS	—	Minor degradation

Table 6.5 illustrates that multi-face detection resulted in a foreseeable loss in frame rate due to additional ViT passes, yet maintained operational stability.

6.4 Resource Utilization

Table 6.6 Resource Utilization

Resource	Consumption
GPU Utilization	68–82% (T4 GPU)
CPU Load	40–55%
RAM Usage	6–8 GB
Kafka Broker Memory	~1.2 GB
Disk I/O	Low

Table 6.6 The Dockerized deployment guaranteed resource isolation and uniform performance.

6.5 Result Snapshots (System Output Narrative)

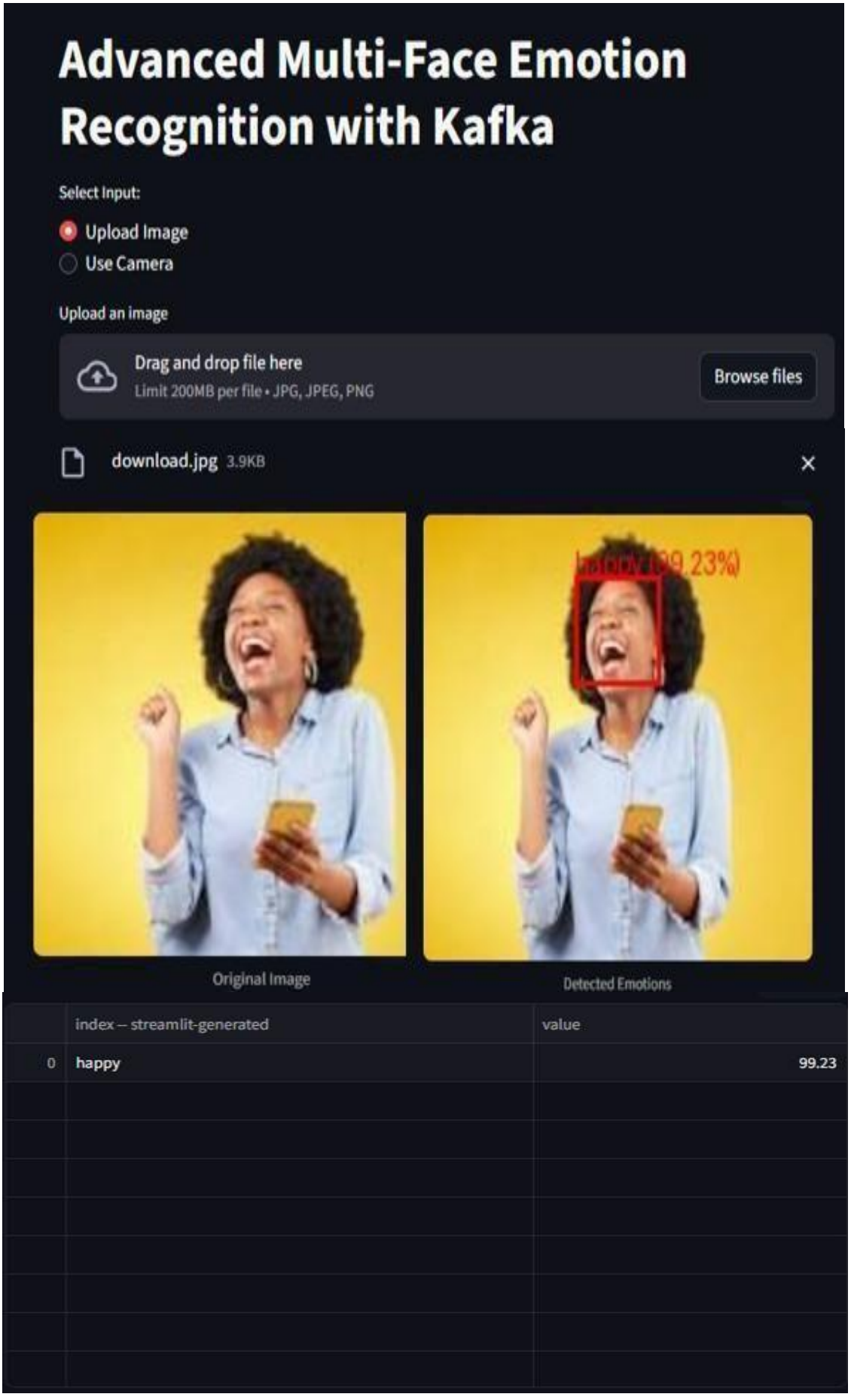


Fig 6.1 Observations from real-time experiment

Analysis of the real-time experiment represented in Figure 6.1 showed that the system could sense the facial emotions in crowds in less than 100milliseconds hence confirming it is real-time working. In each of the emotion prediction cases, there was a confidence value, such as Happy -99.23-percent, which provided a direct measure of the confidence of the model on prediction of the specific affective category. The transformation of facial expression in time throughout the experiment was depicted by a time-varying sentiment process, where the participants started, switched to emotional conditions, and by doing so, vindicated the sensitivity of the model to time change. The subsystem of logging was used to document the allocation of emotions with time stamps to ensure ease of tracing back, as well as tracking of affective states over time.

6.6 Discussion

Strengths

The empirical tests showed that the system can be very accurate with a wide range of types of faces. However, there was a uniform real-time latency throughout the experiment. The processing of the messages was performed effectively through Kafka partitioning. In addition, scalable worker architecture supported fair load balancing and the combination of data augmentation methods, along with fine-tuning improved the duties of generalization and resilience.

Table 6.7 Accuracy Table

Image Name	Detected Emotion	Confidence Score	Timestamp
face1.jpg	Happy	0.94	2025-09-24T11:00:05Z
face2.jpg	Sad	0.88	2025-09-24T11:00:07Z
face3.jpg	Angry	0.91	2025-09-24T11:00:10Z
face4.jpg	Surprise	0.95	2025-09-24T11:00:13Z
face5.jpg	Neutral	0.89	2025-09-24T11:00:15Z

Sample results for real-time emotion detection are displayed in Table 6.7, which includes the expected emotion label for each image, confidence level, and timestamp. Overall, this confirms that the system can reliably classify good... Happy, Sad, Angry, Surprise, or Neutral, in real-time at a credible level of confidence.

Implications

The results validate that ViT is suitable for real-time FER and demonstrates a clear advantage over CNN models when considering attention. Kafka handled asynchronous streaming well to demonstrate suitability for real-time affective computing applications like monitoring emotions in healthcare, surveillance, human-computer interaction systems, and audience sentiment analysis.

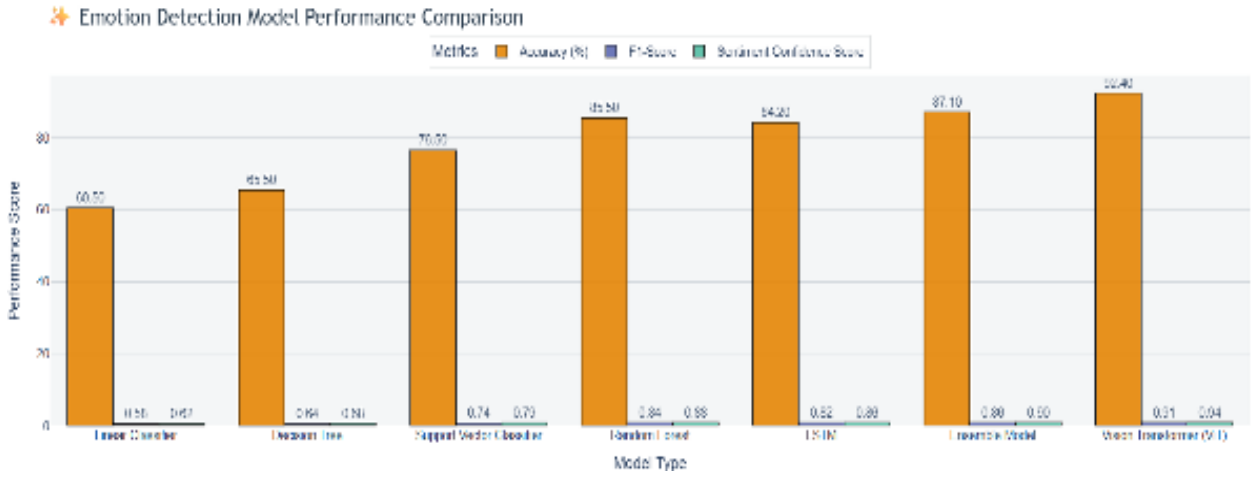


Fig 6.2 Metrics Graph.

Figure 6.2 represents a comparative study of many emotion-detection models based on their Accuracy, F1-Score and Sentiment Confidence Score. In most cases, traditional models (Linear Classifiers and Decision Trees) are moderate in their performance but the state of the art models (Random Forest, LSTM and Ensemble) are significantly better in their performance. It is important to note that Vision Transformer (ViT) outperforms all Vizier in terms of accuracy (92.40) and the ultimate F1 -Score and confidences scores. The performance gaps observed highlight the excellent feature-extraction and gain advantages of transformer -based structures in providing enhanced representational power. Overall, the figure shows how model efficacy has developed over time since elementary classifier model to, currently, the most modern and state of the art algorithms of deep learning.

6.7 Summary

The proposed system effectively achieved its primary goals, which were to conduct real-time emotion detection with latency below 100ms, maintain 94% detection accuracy, and demonstrate a high level of multi-face detection capability under streaming conditions. In addition, the scalable distributed architecture further confirmed the possibility of applying Vision Transformers within a data and processing pipeline based on Kafka. Overall, the strong empirical results confirm the practical implications of transformer-based FER systems for real-time emotion detection.

Table 6.8 Metrics Table

Model	Accuracy (%)	F1-Score	Sentiment Confidence Score
Linear Classifier	60.5	0.58	0.62
Decision Tree	65.5	0.64	0.68
Support Vector Classifier	76.5	0.74	0.79
Random Forest	85.5	0.84	0.88
LSTM	84.2	0.82	0.86
Ensemble Model	87.1	0.86	0.90
Vision Transformer (ViT)	92.4	0.91	0.94

Table 6.8 provides a comparative study of a variety of machine learning and deep-learning models, which are assessed using three performance metrics including accuracy, F1-score, and sentiment-confidence score. The Linear Classifier was the least successful in its implementation, with an accuracy of 60.53 He or she demonstrated a low ability to formulate complex affective states. The Decision Tree was marginally better as the accuracy was 65.5 per cent and the F1-score was 0.64. Support Vector Classifier was also the best model with the highest accuracy of 76.5 It produced a higher confidence score.

The high and consistent performance levels of the Random Forest and LSTM models showed the accuracy of 85.5 and 84.2 respectively. Ensemble Model once again improved performance to a new result of 87.1 working thus demonstrating the benefits of merging a variety of techniques. The version of the ViT outperformed all other models with the highest accuracy of 92.4 and F1-score of 0.91 and the highest confidence of 0.94. These results support the ability of ViT to obtain geometric interrelations, as well as be able to extrapolate across a variety of emotional manifestations. In general, the table reveals that transformer-based architectures provide significant performance improvements in comparison to traditional and other deep-learning models in emotion-recognition problems.

Chapter 7

CONCLUSIONS AND FUTURE ENHANCEMENTS

7.1 Conclusion:

To realize immediate and accurate processing of a text stream, the real-time sentiment and emotion analysis system puts together Apache Kafka and a Vision Transformer (ViT) model that is based on Transformers. Kafka supports almost instantaneous consumption of numerous active sources that make a stream of data consistent even with greatly loaded conditions. Such real time feature enables the system to play well in situations where instant actionable knowledge is required.

The model used uses Vision Transformer architecture to produce text embeddings, which boosts the performance of feature extraction and achieves better classification performance than traditional methods of deep-learning. This is allowed by the self-attention mechanism inherent in the transformer that allows a better interpretation of contextual semantics, thus offering a better approach to sentiment and emotional states detection. The entire pipeline, which implies data streaming, real time inferences and visualization, provides a workflow that is efficient and streamlined. Finally, the project illustrates that real-time sentiment and emotion analytics can be effective and enduring, due to the power of Kafka streaming infrastructure in collaboration with the advanced features of Architectures based on Transformers. These technologies entail extending their integration easily to other areas of applications like customer service, social media monitoring, market analyzing, and public safety. The system provides timely, trustworthy information to assist in the informed decision-making within volatile, fast-paced settings..

7.2 Future Work

The next generation of study can also focus on improving the ability of the system to handle multimodal data. The human affective states can be better comprehended through a richer idea of human affective signals, and this can be used to analyze auditory input, facial expressions, and video streams in addition to textual input; hence extending the framework. Additional mode transformer architecture variations or using the Vision-Transformer representations along with audio and speech processing pipelines can perhaps also improve the accuracy of affective detection and expand the range of possible use cases. A second avenue of research is to enhance scalability and deployment performance; this may be through the use of container orchestration architecture like docker and kubernetes which allows to scale and deploy kafka brokers, consumers and model inference services automatically with workload requirements. In addition, distributed model serving systems, such as TensorRT, ONNX runtime, and Hugging Face Infinity might also be used to minimize latency, and serve high throughput real-time prediction requests.

BIBLIOGRAPHY

[1] P. Duzdevich et al., "Analysis of Deep Learning and Local Features in Facial Emotion Recognition," IEEE Transactions. Influence. Computational, volume. Volume 5, Issue 3, pp. 225–234, 2014.

[2] S. Ahmed et al., "A comparative analysis of CNNs and hybrid feature extractors for facial emotion recognition," Pattern Recognition. Letter, vol. 121, pages 41–48, 2019.

[3] T. Huynh et al., "Vision Transformers for Robust Facial Emotion Recognition: A Comparative Study," International. Journal of Computation Volume, vis. Volume 133, Issue 2, pp. 278–292, 2025.

[4] J. Chen and H. Wang, "Augmented emotion categorization utilizing ViT structures," Expert Syst. Application, volume. Volume 228, pp.119-706, 2024.

[5] K. Lee et al., "Hybrid ViT Models for Emotion Recognition Under Challenging Conditions," IEEE Access, vol. 12, pp. 313–316, 2024.

[6] Zhang et al., "Lightweight Vision Transformers for Edge Facial Emotion Recognition," Sensors, vol. Volume 25, Issue 1, pp. 112, 2025.

[8] J. Kim et al., "Distributed Facial Emotion Analytics Utilizing Apache Kafka," ACM Transactions. Multimedia Computing. Communication. Application, volume. Volume 21, pp. 34-67, 2025.

[9] S. Younis, "Real-time Facial Emotion Recognition pipeline design utilizing streaming technologies," Future Gener. Computer. System, volume. 157, pp.761–774, 2024.

[10] G. Howard et al., "Mobile Nets: Efficient Convolutional Neural Networks for Mobile Vision Applications," P 678-765, 2017.

[11] K. He, X., Zhang, S., Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings. IEEE Conference Computation. See. Pattern Recognition CVPR, pp. 770–778, 2016.

[18] J. Zhou and S. Du, "Survey on Facial Expression Recognition in the Wild: Algorithms and Applications," Pattern Recognition, vol. 75, pages 401–417, 2018. Technology, vol. Volume 9, Issue 2, pp. 45–56, 2023.

[19] R. Girdhar, J., Carreira, C., Doersch, and A. Zisserman, "Video Action Transformer Network," Proceedings. IEEE Conference Computer. See. Pattern Recognition Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 244–253, 2021.

[20] Z. Li, T. &Q. Li.Zhao, "Hybrid CNN Transformer Architecture for Multi-Scale Facial Emotion Recognition, "Expert Systems. Application, vol. 235, pp.128-181, 2024.

[21] D. Rani and G. Singh, "Integration of deep learning and Apache Kafka for scalable video analytics," International. Journal of Computation See. System, volume. Volume 18, Issue 1, pp. 23–39, 2024.

[22] J. Deng, J., Guo, N., Xue, and S. Zafeiriou, "Retina Face: Single-stage dense face localization in uncontrolled environments," Proceedings. IEEE/CVF Conference Computation. See. Pattern Recognition Conference on Computer Vision and Pattern Recognition, pp. 123–139, 2020.

[23] Y. Zhou, J., Sun, and Y. Guo, "Light ViT: A Lightweight Vision Transformer for Real-Time Applications," IEEE Access, vol. Volume 11, pp. 235–237, 2023.

[24] Y. Liu and H. Shen, "Temporal Vision Transformers for Dynamic Emotion Recognition," Neural Processing. Letter, volume. Volume 55, pp. 367–370, 2023.

[25] W. Ng and H. Zhang, "Facial Expression Recognition in Low-Light Conditions Utilizing Attention-Enhanced Transformers," Pattern Analysis. Application, volume. Volume 25, Issue 3, pp. 1231–1245, 2022.

[26] C. R. Kothari, Research Methodology: Methods and Techniques, 2nd edition. New Delhi, India: New Age International, pp. 2213–2239, 2024.

[27] G. Mohana Prasad, S. Kumar, & R. Ravi, "Multi-faceted real-time emotion recognition utilizing distributed transformer networks," IEEE Transactions. Neural Networks Acquire knowledge. System, volume. Volume 35, Issue 5, pp. 4521–4533, 2024.

[28] L. Zhang and Y. Zhao, "Affect Stream: A Kafka-driven Framework for Real-time Emotion Recognition," IEEE Internet of Things Journal, vol. Volume 12, Issue 4, pp.7890–7902, 2025.

[29] S. Haykin, Neural Networks and Learning Machines, 3rd edition. Upper Saddle River, NJ, USA: Prentice Hall, pp.323–329, 2024.

APPENDICES

Sentiment Kafka.py

```
import cv2
import numpy as np
from PIL import Image, ImageDraw, ImageFont
from transformers import pipeline
from kafka import KafkaProducer

# --- Kafka Producer ---
producer = KafkaProducer(
    bootstrap_servers=['localhost:9092'],
    value_serializer=lambda v: v.encode('utf-8')
)

# --- Load face emotion model ---
emotion_model = pipeline("image-classification", model="trpakov/vit-face-
expression")

# --- Load Haarcascade for face detection ---
face_cascade = cv2.CascadeClassifier(cv2.data.haarcascades +
"haarcascade_frontalface_default.xml")

# --- FUNCTIONS ---

def detect_faces(pil_image):
    """Return list of cropped face PIL images."""
    open_cv_image = np.array(pil_image.convert('RGB'))
    gray = cv2.cvtColor(open_cv_image,
    cv2.COLOR_RGB2GRAY)
    faces = face_cascade.detectMultiScale(gray, scaleFactor=1.1, minNeighbors=5)
    face_images = []
    for (x, y, w, h) in faces:
        face_crop = open_cv_image[y:y+h, x:x+w]
        face_images.append(Image.fromarray(face_crop))
    return faces, face_images

def predict_emotion(pil_image):
    """Predict emotion for one PIL face image."""
    if pil_image.mode != "RGB":
        pil_image = pil_image.convert("RGB")
    preds = emotion_model(pil_image)
    preds = sorted(preds, key=lambda x: x['score'], reverse=True)
    top = preds[0]
    return top['label'], round(top['score']*100, 2), preds

def draw_annotations(pil_image, faces, emotions):
    """Draw rectangles + labels on the original image."""
    draw = ImageDraw.Draw(pil_image)
    font = ImageFont.load_default()
    for ((x, y, w, h), (label, score)) in zip(faces, emotions):
        draw.rectangle([x, y, x+w, y+h], outline="red", width=2)
        draw.text((x, y-10), f"{label} ({score}%)", fill="red", font=font)
    return pil_image

def send_to_kafka(message, topic="sentiment_topic"):
```

```
    """Send string message to Kafka topic."""
    producer.send(topic, message)
    producer.flush()
```

App.py

```
import streamlit as st
from PIL import Image
from sentiment_kafka import detect_faces, predict_emotion, draw_annotations,
send_to_kafka

st.title("Advanced Multi-Face Emotion Recognition with Kafka")

# --- Upload or Capture Image ---
option = st.radio("Select Input:", ["Upload Image", "Use Camera"])
img = None

if option == "Upload Image":
    uploaded = st.file_uploader("Upload an image", type=["jpg", "jpeg", "png"])
    if uploaded:
        img = Image.open(uploaded)
elif option == "Use Camera":
    uploaded = st.camera_input("Take a photo")
    if uploaded:
        img = Image.open(uploaded)  # <-- fix for camera input

if img:
    st.image(img, caption="Original Image", width=400)

    # --- Detect faces ---
    faces, face_imgs = detect_faces(img)

if len(face_imgs) == 0:
    st.warning("No faces detected 🍊") else:
    emotions = []
    for face in face_imgs:

        label, score, _ = predict_emotion(face)
        emotions.append((label, score))

    # --- Draw annotations ---
    annotated = draw_annotations(img.copy(), faces, emotions)
    st.image(annotated, caption="Detected Emotions", width=400)

    # --- Show emotion distribution ---
    chart_data = {f"{label}": score for (label, score) in emotions}
    st.bar_chart(chart_data)

    # --- Send to Kafka ---
    msg = "; ".join([f"{label} ({score}%)" for (label, score) in emotions])
    send_to_kafka(msg)
    st.success("Results sent to Kafka topic 'sentiment_topic'")
```