CUSTOMER EMOTION DETECTION FOR IN-STORE EXPERIENCE

ABSTRACT

Facial emotion recognition is at the forefront of improving in-store customer experience through the ability to decipher actual real-time emotional reactions. However, despite the advancements in deep learning, current models tend to remain deficient in balancing spatial feature extraction with contextual interpretation of facial expressions. This research suggests a detailed comparative examination of three DL models: (1) CNN + Transformer + EfficientNet, (2) CNN + Transformer, and (3) FaceNet_CNN, for strong facial emotion recognition. Implemented on TensorFlow and Keras, models are trained and tested against the FER-2013 dataset consisting of more than 35,000 facial images labeled with seven emotions. Grayscale normalization, face alignment, resizing, and augmentation are preprocessing steps. The best is achieved by the hybrid model coupled with EfficientNet at 74.8%, followed by the CNN + Transformer model (71.2%) and FaceNet_CNN (69.3%). Spatial fine-grained features are effectively extracted by the EfficientNet module, and long-range dependencies of facial landmarks are learned by the Transformer. FaceNet_CNN, a robust baseline for facial identity recognition, falls behind in emotion classification as it has limited contextual modeling. This performance improvement—a 5.5% increase over baseline—demonstrates the synergy of the combination of local and global feature extractors. The suggested approach provides a real-time and scalable solution that can be incorporated into retail analytics platforms with TensorRT or ONNX for optimized inference. Through overcoming the limitations of previous works and utilizing advanced architectures, this research provides a basis for more accurate, adaptive, and commercially viable emotion recogntion systems in customer-oriented settings.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

Hybrid CNN + Transformer

Hybrid CNN + Transformer + EfficientNet

FaceNet_CNN

LIST OF IMPORTANT ABBREVIATIONS

Convolutional Neural Network (CNN)

Facial Emotion Recognition (FER)

Region of Interest (ROI)

Natural Language Processing (NLP)

Artficial Intelligence (AI)

Machine Learning (ML)

Deep Learning (DL)

Residual Network (ResNet)

Deep Neural Network (DNN)

Rectified Linear Unit (ReLU)

CHAPTER 1- INTRODUCTION

In modern retailing, enhancing the customer experience is one of the major strategies employed by businesses looking to stay ahead in the competition. More traditional methods of customer satisfaction measurement, such as comment cards and surveys, are too focused and do not register emotional feedback in real time (Khomidov and Lee, 2024). Given the latest advancements in AI and computer vision, emotion recognition is now an available choice to bridge this gap. By reading and interpreting the customer's emotions via facial expressions, body language, voice tone, or physiological indicators, customers can be understood better by retailers as they go through their shopping experience.Emotion detection uses technologies like CNNs, Transformer-based models, and extremely powerful models like EfficientNet to analyze multimodal inputs (Sumon *et al.*, 2025). These allow stores to react in real time to consumers' emotional state, customize products, and ultimately influence engagement and satisfaction. For example, frustration triggers may lead to support in real time, and interest moments can be utilized for product recommendations. These emotionally intelligent platforms create the retail environment a responsive and human touch-driven one. (Dalvi *et al.*, 2021).

The need for real-time analysis and interactive customer service has driven research towards creating models that not only yield correct results but are also computationally efficient to apply in retail settings.EfficientNet, with its adaptive architecture, has performed well in resource-constrained environments (Khomidov and Lee, 2024), while Transformers enhance the ability to handle contextual and temporal emotional information. When combined with CNNs as a feature extraction module, they constitute a robust system for emotion detection that is capable of handling difficult audio and visual inputs. The combination of these technologies facilitates the creation of multimodal emotion recognition systems that can

process visual, audio, and text inputs to offer a comprehensive view of customer emotions. For instance, when a customer shows signs of frustration or confusion, a system with these models can trigger staff intervention or provide support through smart displays (Nguyen and Mogaji, 2023). Likewise, positive sentiments such as delight or surprise may be utilized for customized promotions or rewards for loyalty, improving the possibility of conversion and return visits.

The customer experience is difficult for retailers since it is determined by factors which are within the control of the retailer and factors that are outside their control. Most of the conventional traditional supermarkets have lost business to these new store formats. In summary, the large new supermarket rivals belong to three phenomena. There are Makro and Game that sell food and groceries and they have Walmart's negotiating power that enable them to be more price aggressive. There are secondly food specialist shops. The competitive nature of the food forms that new foods introduce is that they do not carry as high a stock keeping units (SKU's) as supermarkets and focus on fewer food SKU's with more turnover for profit. In the third instance, we have convenience stores that stock a larger merchandise mix of profitably consumed high-frequency items (Terblanche, 2018).

By examining facial cues recorded through webcam, the system is able to forecast user mood with high reliability, generating an adaptive virtual shopping space that conforms dynamically to the mood of users (Bandyopadhyay, Thakur and Mandal, 2024a). In this research, we present a multimodal emotion recognition model that merges EfficientNet, CNN, and Transformer models, CNN and Transformer and FaceNet_CNN models to detect customer emotions in real-time in-store environments. The proposed model is balanced with respect to accuracy, interpretability, and computational expense. We will examine the methods for hyperparameter tuning to achieve maximum performance and contrast the results with existing state-of-the-art models. The goal is to develop an emotion-aware system that

enhances customer experience, enhances business decision-making, and promotes intelligent retail automation.

Emotion detection is not just a technology breakthrough it's a game-changing tool that harmonizes the emotional intelligence of in-store settings with human-to-human interactions. When implemented with good model architecture and ethics, emotion detection can change the way businesses know and serve customers.

1.1 Background of the Study

In retail, emotional intelligence of the customers has become a powerful tool to improve the shopping experience, decide purchasing behavior, and increase customer satisfaction. Emotion detection covers various technologies and methods for detecting and analyzing customers' emotional states in real-time. Since consumer behavior is increasingly being driven by individual experiences, emotional intelligence within retail has become a prime driver in influencing customer interactions and improving in-store experiences.

Facial Expression Recognition (FER) is widely utilized as one of the most popular methods of emotion detection among customers. It is based on facial muscle movement analysis to determine emotions like happiness, anger, sadness, or frustration. Facial expressions are a cross-cultural non-verbal cue, and thus FER is an extremely useful method of emotion detection. FER-based retail systems assist owners and customer care agents in determining customer satisfaction or anger and real-time adjusting the setting or service. Facial emotion recognition has enhanced customer service a great deal by offering feedback about a customer's emotional status, allowing sales attendants to personalize interactions and raise the success rates of transactions, as asserted by (Jia *et al.*, 2021)

The significance of emotion detection in retail extends far beyond maximizing customer satisfaction it also gives companies the power to optimize strategies in real-time to allow customers to feel valued and engaged. With emotion-aware technology continuing to grow, it holds the promise to revolutionize the way retailers are thinking about customer service and store design.

1.2 Significance of Emotion Detection in Retail

Customer emotional understanding has become a key retail strategy to improve the store experience, especially in brick-and-mortar stores. Standard retail measurements like sales metrics, dwell time, or customer demographics tend to be insufficient to record shoppers' actual emotional status in real time.Emotion detection, powered by AI and computer vision advancements, provides an invaluable glimpse of customers' emotions in the form of facial expressions, body posture, tone of voice, and physiological cues (El Ayadi, Kamel and Karray, 2011).Emotionally intelligent shops have the ability to personalize experiences and improve satisfaction.For instance, emotion recognition systems can signal employees to approach an angry customer or suggest products to a smiling, engaged shopper. Studies prove that emotion-sensitive systems in retailing improve customer retention and build stronger brand loyalty (Lu *et al.*, 2023). Through the adaptation of in-store atmospheres lighting, music, layout according to real-time emotional responses, retailers are able to make passive shopping experiences into interactive, customized ones.

Current research points out that real-time emotion detection also enables customer segmentation and predictive analytics Emotion recognition model using DL that operates on facial expression analysis to enable product recommendation systems (Bandyopadhyay, Thakur and Mandal, 2024a). While their application was focused

on online platforms, the approach and architecture using CNNs for feature extraction from images and transformers to learn sequential patterns are equally applicable to brick-and-mortar stores. The integration of these models results in a high-accuracy, scalable real-time shopper emotion monitoring and responding system. Additionally, emotion detection in brick and mortar stores can be beneficial for operations. Store layouts can be optimized to ensure maximum by tracking emotional response to product placements or promotions. This data driven strategy enhances marketing campaigns, decreases churn, and maximizes overall store performance.

Additionally, emotion detection is crucial in staff training and performance assessment. AI systems can evaluate customer interactions and give feedback on employee conduct based on emotions detected. This promotes a customer-focused culture in the organization and improves service quality in the long run. In spite of its promise, incorporating emotion detection in retail also has ethical and privacy issues. Consent, data storage, and transparency have to be dealt with carefully in order to remain compliant with data protection laws such as GDPR. Nevertheless, if properly implemented, emotion recognition technology provides a competitive edge to retailers in their attempts to humanize their service and make a lasting impression. Overall, emotion detection in retail is not just a technological upgrade it's a strategic tool for making empathetic, personalized, and data-driven retail experiences. As shopping in stores goes on changing in a post-pandemic world, such technologies have the potential to reshape customer interaction and business performance.

1.3. Research Problem

The focus of this research is to create a real-time emotion detection system of customers that builds an improved retail experience within stores by adjusting

services according to emotional indicators. The research question is on the identification of the best hybrid DL method for the efficient and accurate classification of emotions in stores.

Research Question: How well can customer emotions be detected through the use of deep learning techniques based on facial expressions in stores?

Aim: To recognize and classify customer emotions in physical stores using facial expressions by comparing and evaluating the performance of three deep learning models—CNN + Transformer + EfficientNet, CNN + Transformer, and FaceNet_CNN—to identify the most efficient and accurate model for real-time use.

Objectives of the Study: To demonstrate the performance gain by combining Transformer's contextual characteristics with CNN's spatial feature extraction compared to traditional or single-architecture models

Hypothesis: The CNN + Transformer model will perform better than the CNN + Transformer + EfficientNet and FaceNet_CNN models both in emotion classification accuracy and real-time processing efficiency. This is because it achieves the best possible balance between feature extraction (CNN) and temporal/contextual learning (Transformer), making it the best approach to be deployed in emotion-sensitive, real-time retail environments.

1.4. Scope of the Study

- To formulate a hybrid model of emotion detection using EfficientNet, Transformer, and CNN architectures to recognize facial emotions in real time in retail scenarios.

- To utilize the suggested model on real or captured video feeds of in-store CCTV cameras to recognize customer emotions including happiness, sadness, anger, fear, and surprise.
- To compare the model's success against metrics such as exactness, processing time and real-time utility in in-store environments.
- To examine the differences in customer emotional reactions in different shopping contexts, such as browsing products, waiting in queues, and customer service interactions.
- To offer suggestions for store managers and business analysts to leverage emotion data to tailor shopping experiences and engage customers.
- To help implement emotion-aware business strategies like adaptive digital signage, mood-based lighting or music, and real-time employee support.
- To investigate how customer emotion data can be applied to optimize store layout, targeted promotions, and customer retention programs.
- To provide an emotion detection framework based on DL that may be extended in further research to add voice tone, body posture, or multimodal emotion detection**.**
- To follow privacy and ethical guidelines, with the reminder that data collection and processing have to be in accordance with consumer protection policies.
- To limit the study to emotion detection using facial expressions, with no mention of physiological or audio based ones in this work.

1.5. Limitations

This research considers only facial expression-based emotion recognition and does not involve multimodal cues such as voice, body position, or physiological signals. The model can perform sub-optimally in busy or low-light retail settings and can only recognize basic

emotions and possibly miss out on complex or subtle emotional conditions. Cultural and demographic variations in facial expressions can affect accuracy due to constraints in datasets. Real-time deployment would be constrained by hardware requirements, and observable cooperative customer behavior is assumed by the system, which may not always be a valid assumption. Privacy issues and ethical considerations are mentioned but not deep-dived into, and integration with retail management systems is out of scope for the project at the moment. In addition, the subjective nature of emotion ensures that expressions of identified sentiments are not nesscessarily indicative of true customer satisfaction or intent.

1.6. Roadmap

Figure 1.1 presents the outline of the dissertion

Figure 1. 1Roadmap of research

In later section i.e., section 2, we have talked more about the earlier work done on this topic of research, which will then be followed by the discussion on methodology employedfor the implementation of our model in section 3. The discussion of results is done in section 4, wherein we test the performance of the model. Section 5 includes the conclusion of the research.

CHAPTER 2 – LITERATURE REVIEW

Recent progresses in affective computing have made it possible for vendors to better understand and respond to customers' emotions in store shopping settings. The literature discussed describes how affective responses elicited by facial expressions have the key role of shaping customer experiences and informing consumption behaviour. The researches in these studies recognize the role played by in-store factors such as ambient, contact with employees, and product placement towards emotional engagement. Various methodologies have been utilized, from deep learning models and real-time video analysis to sensor-based tracking and zero-shot learning algorithms. Techniques applied for raising accuracy in detecting emotion include SNNs, CNNs, and optimization algorithms. While there are difficulties like data privacy, hardware constraints, and generalizability, the overall perception is in favor of the importance of emotion detection systems in creating more personalized and responsive shopping environments.

Bandyopadhyay,S (2024b) developed an emotion-based real-time online recommendation system based on facial expression analysis. The process is to record users’ facial expressions by a webcam while shopping online and analyse them by a deep learning model that is trained on facial image databases. Emotions are identified from five main facial expressions, and recommendations are made dynamically without any reference to historical ratings or purchase history. The system had a real-time emotion recognition accuracy of 75% and produced product recommendations accordingly. Customer feedback was used to test its performance. The method provides the benefit of personalization and adoptability in recommendations, making virtual shopping more enjoyable. It has drawbacks such as medium accuracy, webcam input dependence and possible privacy issues.

Nguyen et al.(2022) proposed an AI-driven model to enhance store design layout optimization using existing CCTV infrastructure to discover and enhance customers' in-store experience. The methodology involves thorough analysis of current store layout methodology, followed by integration of modern AI trends i.e., computer vision and deep learning to analyse videos recorded from surveillance networks. This system is designed to monitor customer navigation patterns, dwell time throughout sections, and emotional responses to in-store merchandising. While the study itself is a conceptual proposal, the system itself is designed to parse actual real-world video feeds from CCTV cameras in order to produce actionable insights. The desired end result is data-driven, adaptive store layout design that is closer to actual customer behaviour than it would be using sales data alone. The input data for this approach would be primarily in-store video monitoring with movement path and emotional expression annotations. The advantages are real-time layout optimization, increased sales opportunity, and increased customer satisfaction through tailored experiences. The disadvantages are ethical concerns related to surveillance, complexity in emotion detection in varied situations, and the challenge of integrating such systems into existing store operations without violating customer privacy or store processes.

Mellouk and Handouzi (2020) developed an recent advancements in automatic facial emotion recognition (FER) with deep learning methods. The main aim is to investigate the progress from 2016 to 2019 in using DL architectures, including CNNs and RNNs, for emotion recognition from facial expressions. The paper contains discussion of some of the popular open access datasets like FER2013, CK+, and JAFFE which are used frequently for training and testing of FER models. The authors highlight the benefits of deep learning methods over conventional approaches, most notably their capacity to learn characteristics automatically and enhance classification accuracy without human intervention. The article functions as an essential reference guide to researchers through a comparison of performances by the models,

an outline of architectural innovations, and directions for future development. Yet the research also highlights some challenges embedded in FER, such as differences in pose, illumination, age, gender, background, and occlusion due to adornments like scarves or spectacles. Also, the article recognizes that any model cannot remain stable in every condition and in all groups. Another limitation pointed out is the enormous computational power and vast data requirements required to train deep learning models. In general, the review emphasizes the advancements made in FER using deep learning but highlights major areas where research is critical.

Gaddam et al.(2022) proposed a facial emotion recognition system using deep learning, especially applying the ResNet50 architecture in human emotion classification using static images. The model is learned over the FER2013 dataset comprising image-annotated basic human emotions such as happiness, sadness, anger, and surprise. Deep CNN provides for automatic feature learning, resulting in far improved classification accuracy than previous machine learning models. The study highlights CNN's improved performance in visual information processing and achieving robustness in emotion classification tasks. The most significant strength is ResNet50's ability to effectively process complex image features. However, the model's reliance on static images and one dataset presents limitations towards generalization in real-world scenarios and flexibility across different conditions.

Yolcu et al. (2020) designed a system which handles the assessment of customer interest through the analysis of facial expressions and facial orientations. The approach includes separating frontal face poses, partitioning important facial parts to create an iconized face image, and further analysing facial expressions through the fusion of features from the iconized and raw images. This approach combines part-based local features and holistic facial information to enhance recognition performance. While the specific dataset used isn't named in the abstract, strengths of the system include its ability to be used with low-cost imaging

equipment like webcams and potential application to tracking emotional responses in focus groups. But limitations could comprise challenges in accurately interpreting ambiguous or subtle facial expressions and privacy issues of concern with the collection of facial data. Experimental results verify excellent accuracy in identifying customer interest, making it ideal for intelligent retail and human-computer interaction applications.

Pantano (2020) investigate the potential for the systematic assessment of retail service encounters by identifying consumers' emotions via facial expressions. The technique aims to assist retail staff in comprehending customers' experiences more effectively with the aid of automated mechanisms. The suggested methodology uses a machine learning-based system capable of identifying six basic human emotions like happiness, anger, fear, disgust, surprise and sadness through non-verbal facial cues. The model was validated through facial data gathered from consumers in the 19 biggest shopping centers of the UK, under actual shopping experiences. Results show that facial expression of retail service encounters. Additionally, results indicate general consumer acceptance for such systems towards automatic judgement purposes. Benefits of the method are the capacity to give real-time, non-intrusive emotional feedback, raise employee awareness, and maximize real-time customer service. Drawbacks are ethical issues involving privacy or consent, possible misclassification of emotion based on cultural or contextual differences and dependence on ideal lighting conditions and camera angle for accurate emotion classification.

Wen, Abe and Suganuma (2022) constructed an adaptive and effective Customer Behavior Recognition (CBR) system that can handle newly occurring or evolving behaviors within shopping environments without requiring retraining. The novel approach employs a primitive-based recognition mechanism, under which behaviors are specified as composites of simple action primitives. Support Vector Machine (SVM) and Random Forest (RF) machine learning algorithms are utilized for classification of these behaviors. The system was

validated based on both real-world and simulated datasets that comprised video data from stores with observations of diverse customer activities. Results indicate that the approach has high accuracy in recognizing behavior and exhibits good adaptability to novel behavioral patterns. Among its major strengths are support for integration into existing CCTV infrastructure, effectiveness in identifying broad categories of customer behavior, and flexibility to add new behaviors without having to retrain the entire model. The method does demand manual definition of primitives and could be limited in dealing with more subtle or complex customer behaviors. Also, since the system does not employ deep learning, it might be deprived of the advantage of abstraction at high feature levels normally provided by deep neural networks.

Mazhar et al.(2022) illustrated how facial recognition video analysis can be utilized for sentiment analysis to aid business growth and decision-making. The suggested approach is a lightweight machine learning algorithm for aspect-oriented emotion classification from movie reviews. Emotions are identified through facial expressions in video data, processed through machine learning models such as Naive Bayes, SVM, Random Forest, and CNN.The study uses real and publicly available data to evaluate performance. There is high effectiveness with 84.72% accuracy, 79.24% sensitivity, 90.64% specificity, and 90.2% precision. The strengths are enhanced emotion detection accuracy and usage on real-world video-based data. However, noisy data management, dynamic facial variations, and real-time processing are challenges.

Khare et al.(2024) aimed to give a detailed account of emotion recognition methodologies evolved over the last decade and how Artificial Intelligence (AI) specifically machine learning and deep learning has spurred the field. A systematic review of the literature involving reading numerous published studies to discover current practices, popular algorithms, and challenges of emotion recognition has been done by the authors. The paper

reviews several AI-based models that have been used to apply human emotional recognition from facial, speech, and physiological information. It also mentions widely used emotion datasets like FER2013, CK+, and EMODB that are crucial to train and test the recognition systems. The main contribution is that deep learning techniques, namely CNNs and RNNs, have made a huge difference in the accuracy of emotion recognition, making real-time applications more feasible. The main strength of the paper is that it identifies trends and areas of research, and can act as a good guide to future research. But it also outlines limitations such as dataset imbalances, lack of generalizability across populations, and ethical issues of privacy and consent. The paper focuses extensively on the deployment of multimodal methods and explainable AI to guarantee reliability and transparency.

Kusal et al. (2021) proposed an AI-powered emotion recognition in text big data from Online Social Media (OSM) to bridge the increasing demand for fine-grained emotion understanding over sentiment analysis. The approach comprises qualitative inspection of emotion models, datasets, and algorithms, as well as quantitative bibliometric examination based on 910 Scopus and Web of Science papers (2005–2020). The findings underscore prominent contributions, trends in research, and demographic information in this area. Data sources are diverse social media forums such as Twitter and Reddit. Strengths are scalable emotion comprehension over areas like education, healthcare, and human–computer interaction. Weaknesses are the management of unstructured large-scale data, heterogeneous emotional expressions, and non-standardization in emotion models.

Sarvakar et al. (2023) proposes a new approach to facial emotion detection using the Facial Emotion Recognition using CNN (FERC) system. The FERC system is organized into two components: a CNN for feature extraction and an emotion classifier for emotion classification. It uses a CNN framework to automatically learn and feature extract from facial images and then a classification layer for emotion identification. It is tested using standard

facial emotion data sets, proving to be more accurate compared to conventional methods. The FERC system benefits by automatically learning meaningful features without human intervention and with better performance in emotion recognition tasks. Yet, limitations could include that it would require large computational power to train the CNN and possible difficulties in generalizing across various facial expressions. The research contributes to affective computing through providing an efficient method for facial emotion recognition using deep learning methods.

Mehendale (2020) developed a strong facial emotion recognition system with the help of CNN, known as Facial Emotion Recognition using CNN (FERC). The suggested approach involves a two-stage CNN model, wherein the first stage of removing the background of facial images eliminates unnecessary features and the second stage extracts a 24-dimensional expressional vector (EV) for classifying emotions. The model aims at five principal emotional states: anger, sadness, happiness, fear, and surprise. Large-scale experimentation was undertaken using over 750,000 face images obtained from benchmarking datasets such as the Extended Cohn–Kanade (CK+), Caltech Faces, CMU, and NIST. The FERC system recorded a high accuracy of classification of 96%, proving its efficiency in detecting facial emotion. With the involvement of background removal before feature extraction, the model effectively removes noise and enhances the CNN training process, presenting an advantage over traditional single-stage CNN approaches. However, some limitations exist, including the attention to only five fundamental emotions and possible difficulties in real-world settings with occlusions or varied lighting conditions.

Akhand et al.(2021) suggests the model that helps to improve the FER accuracy using transfer learning methods with DCNNs. Here, the approach is to employ a pre-trained DCNN model and re-configure its dense higher layers to the FER task and later fine-tune it on facial emotion datasets. This method leverages transfer learning in effectively transferring

established DCNN architectures for better FER performance. The authors demonstrate that their model has good recognition accuracy, confirming the efficacy of transfer learning for FER tasks. The benefits of this method are less training time and the capability to use pre-trained models, thus avoiding the necessity of large labeled datasets. However, limitations are that restrictions may include potential overfitting in the case that the pre-trained model is poorly fine-tuned, and the success of the method relies on the suitability and quality of the pre-trained model to the FER task. The work contributes to affective computing through showing an example of practical use of transfer learning for enhancing FER systems.

Rana et al.(2024) developed an emotion recognition model that is able to classify and identify customer emotions from facial expressions in real-time. The study entails capturing post-shopping facial pictures of customers and applying machine learning algorithms to analyse and categorize the emotional states depicted on these faces. The study employs a facial image dataset taken from participants subsequent to their shopping experience. The findings show that the suggested model efficiently detects an array of emotions, offering implications regarding customer satisfaction and shopping experience. The merits of this practice are that the potential for retail businesses to be provided with immediate feedback on the customer experience makes it possible for them to upgrade the quality of services and offer personalization. However, limitations are customer privacy concerns and the need for big and diverse datasets to maximize model accuracy and generalizability. The study contributes to the field of consumer behaviour analysis as it offers a novel approach to measuring customer emotions through facial expression recognition.

Table 2.1 This table provides a brief overview of key studies on customer emotion detection in retail, summarizing their purpose, methods, benefits, and limitations across various technological approaches.

Authors	Source	Keywords	Findings
Bandyopadhyay,S.(2024b)	Springer	Recommender systems, deep learning	Proposed a deep model for personalized suggestions.
Nguyen et al.(2022)	Springer	Store layout, AI, smart retail	Reviewed the integration of AI in optimizing physical retail space layouts.
Mellouk and Handouzi (2020)	Elsevier	Deep learning, FER, review	Summarized architectures and datasets used in facial emotion recognition.
Gaddam et al.(2022)	Springer	Deep learning, facial expressions	Showcased a deep CNN-based architecture for facial emotion recognition.
Yolcu et al. (2020)	Springer	Face analysis, customer engagement, DL	Introduced a deep learning system to monitor and interpret customer interest in retail spaces.
Pantano (2020)	Elsevier	Facial expressions, retail services	Used facial emotion cues to assess customer reactions to service experiences.
Wen, Abe and Suganuma (2022)	MDPI	Behavior recognition, adaptability, sensors	Developed a flexible emotion-behavior detection system adaptable to varying customer types.
Mazhar et al.(2022)	MDPI	Movie reviews, face recognition, emotion	Merged textual reviews with image-based emotion detection.
Khare et al.(2024)	Elsevier	AI, emotion recognition, review	Reviewed trends, gaps, and future scope in emotion detection via AI.
Sarvakar et al. (2023)	Elsevier	CNN, FER, image processing	Demonstrated CNN-based method for recognizing facial emotions with training on visual data.
Mehendale (2020)	Springer	CNN, FER, deep learning	Introduced a CNN model tailored for facial emotion classification.
Akhand et al.(2021)	MDPI	CNN, transfer learning, face recognition	Improved accuracy via deep CNN + transfer learning.
Rana et al.(2024)	Springer	Facial analysis, retail feedback, emotion.	Investigated consumer emotional states after shopping using face-based analytics.

Table 2. 1 Summary of Literature reviews

Despite numerous developments in in-store customer emotion detection, there are still some research gaps in current work. One of the biggest limitations is that most models have not been tested in real-world applications since they are typically tested in controlled or simulated environments, which makes them less useful in practice. Many of the models are also limited to recognizing just primary affect states like happy, sad, or neutral lacking the complexity and richness of human affect. Small and uniform data sets with restricted demographic representation also limit the ability to generalise results. Several systems display environmental sensitivity so that performance, especially, remains greatly dependent upon aspects such as illumination and location of sensors. Real-time monitoring of customers in ethical and privacy matters is scarcely given attention. Furthermore, experiments largely use a single-modal input, while bringing together multimodal information integrating facial expressions may achieve greater accuracy. Computational complexity and the requirement of high processing power restrict the scalability of optimized models such as those employing metaheuristic algorithms or spiking neural networks. Furthermore, studies involving special groups such as blind shoppers are not common, and this points towards a lack of inclusivity. Longitudinal studies that look at changes in emotional behavior across time are also lacking. Lastly, while emotional states are detected, few studies examine how the detected emotions actually impact consumer choice, leaving an essential gap between emotion detection and effective retail action.

The literature on emotion detection emphasizes the growing application of sophisticated technologies like facial recognition, EEG signals, speech analysis, and machine learning models to detect customer emotions in retail settings. Past research has shown the potential of these methods to enhance customer experience, personalize services, and enhance store operations. However, most such systems exist within controlled environments only, support limited classes of emotion, and do not always provide real-time functionality or integration

into in-store actions. Furthermore, data privacy concerns, hardware limitation, and lack of inclusiveness issues remain pervasively contentious issues. For these shortcomings to be addressed, the present study aims to advance a real-time, in-store emotion sensing system based on facial expression analysis. The system will yield actionable customer satisfaction measures that enable retailers to react and improve shopping experiences.

CHAPTER 3- METHODOLOGY

3.1. Introduction

The current research proposes a deep learning-based hybrid framework for customer emotion recognition in retail environments and compares three combinations of models: CNN + Transformer + EfficientNet, CNN + Transformer, and FaceNet_CNN. The primary goal is to design an effective system capable of identifying customers' emotional states from facial expressions collected via in-store cameras, thereby enabling real-time measurement of experience for retailers. All methods leverage intrinsic strengths EfficientNet for extracting high-level features CNNs for the detection of localized spatial features, Transformers for the capture of long-range dependencies, and FaceNet_CNN for facial feature embedding. The model follows a standard pattern of data acquisition, preprocessing, model building, training, testing, and deployment. The FER-2013 dataset is used as the main repository for training and testing, which offers a general benchmark for facial emotion classification. This chapter describes the overall methodology adopted in developing and testing the proposed emotion recognition systems.

3.2. Dataset Description and Analysis

The FER2013 dataset is used as the baseline for training and testing our emotion detection model. The dataset is used because it is holistic in nature and highly similar to facial emotion recognition tasksIt contains 35,887 gray-scale facial images of the size 48×48 pixels partitioned into five emotion classes: anger, happiness, sadness, surprise, and neutral. The information is officially categorized into three categories:

- Training set: 28,709 (80%)
- Validation set: 3,589 (10%)

- Test set: 3,589 (10%)

The dataset features a number of challenges typical for real-world facial emotion recognition:

- Lighting variability
- Variation in facial orientations and poses
- Occlusions (glasses, hair, hands)
- Facial feature variation across different population groups
- Unclear expressions potentially belonging to several emotion classes

These challenges reflect very closely the ones faced by retail environments, which makes the FER2013 dataset most appropriate for our task while requiring strong preprocessing and augmentation techniques.

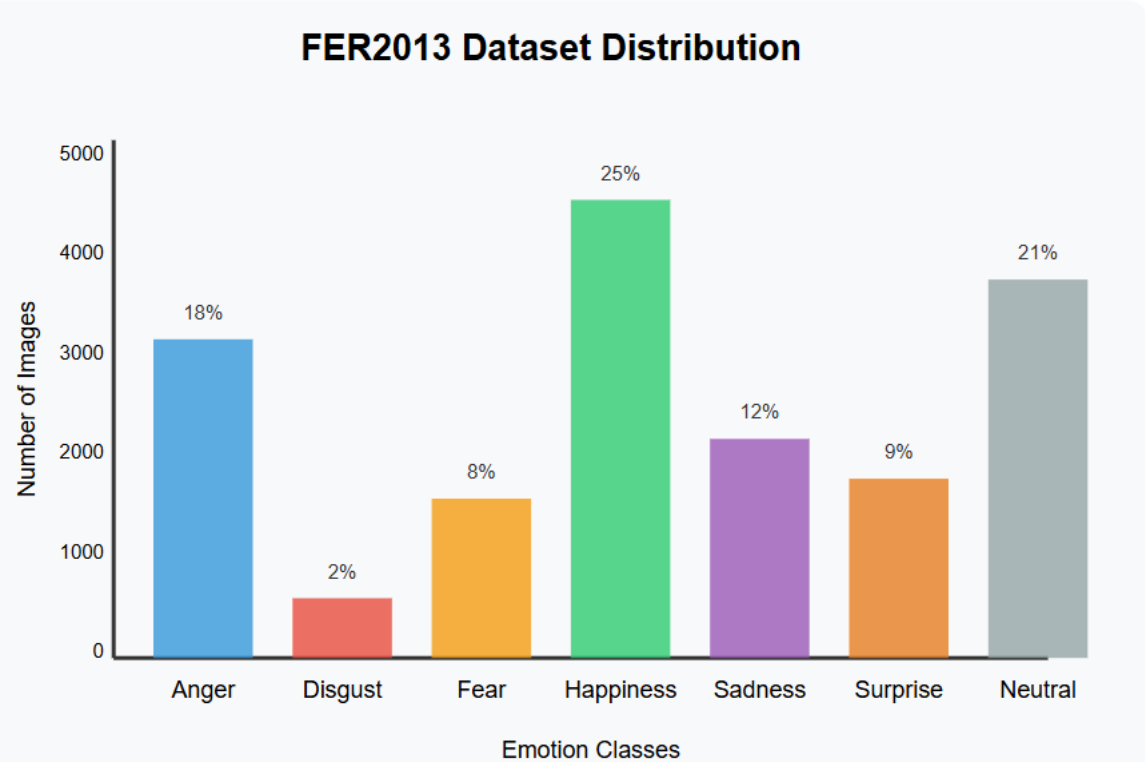

Figure 3. 1Emotion Class Distribution in FER-2013 Dataset

Figure 3.1.shows the spread of images over seven emotion classes in the FER-2013 database. "Happiness" is represented most at 25%, and "Disgust" is represented least at just 2%, which illustrates extreme class imbalance.

3.3. Data Preprocessing

Facial emotion recognition is a difficult task comprising recognizing and decoding human facial expression and labeling it as one among discrete categories of emotions. Most popularly employed data for performing this task includes the FER-2013 dataset, giving rise to training and testing platform for a plethora of deep networks. Although various models like CNN + Transformer + EfficientNet, CNN + Transformer, and FaceNet_CNN have specific architectural needs, they all tend to follow a common set of basic data preprocessing operations. These operations clean the data, make it consistent, and prepare it for feeding into neural network models.

3.3.1. Loading and Parsing the Dataset

First, load the FER-2013 dataset, which is typically provided in CSV format. The dataset rows contain three significant components: the emotion label, a string of pixel values, and a usage type (training, public test, or private test). The pixel values are a flattened 48x48 grayscale image presented as a single string. The initial preprocessing step is to parse the pixel strings and reshape them into 2D arrays in preparation for reconstructing the images. This is performed to transform raw data into a structured form that may be processed further. After conversion, each image is a 48x48 grayscale image of ahuman face that shows one of five emotions: Angry, Happy, Sad, Surprise, and Neutral. To transform the string into a matrix it is obtained in eqn. (1)

$$I(i,j) = reshape(p, 48 \times 48) \quad (1)$$

Where $I(i,j)$ is the grayscale image matrix, $p = [p1, p2, p2304$ is the pixel vector.

3.3.2. Grayscale to RGB Conversion

Though the FER-2013 images are inherently grayscale (i.e., single-channel), the majority of pre-trained deep learning models like EfficientNet and FaceNet are implicitly three-channel RGB-image-capable and trained on the ImageNet dataset. For compatibility with these models, one should convert the grayscale images to pseudo-RGB format.

This conversion is done by tripling the single grayscale channel three times to mimic an RGB image. This doesn't alter what is within the image, but it brings the input format in line with standard deep learning models' assumptions. This operation is particularly necessary in transfer learning scenarios, where the application of pre-trained weights of networks that were trained on RGB can lead to a significant amount of improvement in model performance.

$$I_{RGB}(a,b) = [I(a,b), I(a,b), I(a,b)] \quad (2)$$

Where $I(a,b)$ is the grayscale pixel at position (x,y), $I_{RGB}(x,y)$ is the replicated RGB pixel.

3.3.3. Resizing the image

After getting the images in RGB, the second thing to do is resize them to comply with the input size specifications of the target model (He *et al.*, 2016). Various models require different input sizes. EfficientNet, for instance, tends to require input sizes of 224x224 pixels, while FaceNet tends to require 160x160 pixels. Hybrid models or bespoke CNNs may use input sizes such as 48x48 or 96x96, as per their architecture. Resizing is achieved via interpolation techniques that rescale the image size but attempt to preserve key facial features. Resizing is a crucial step because inputting images of different sizes into a model may cause shape mismatches and model accuracy degradation. The resized image $I_{resized}$ is derived in eqn. (3)

$$I_{resized} = resize(I_{RGB}, X', Y') \quad (3)$$

Where X', Y' are the target height and width respectively.

3.3.4. Normalization of Pixel Values

Image normalization is one of the most important preprocessing steps in deep learning workflows Normalizes the pixel value range to ensure strong and stable model training. Pixel values for the FER-2013 dataset range between 0 and 255(Ioffe and Szegedy, 2015). Some normalization techniques may be utilized based on how the model will be trained.

For basic CNNs, normalization is usually done by scaling pixel values to a [0, 1] range by dividing by 255. This method simplifies computational complexity and speeds up convergence during training. On the other hand, models such as FaceNet involve normalization to the range [-1, 1], where 127.5 is subtracted from each pixel and then divided by 128.0. This method brings data around zero and enhances performance if used with certain activation functions.

EfficientNet, being based on ImageNet pre-trained weights, needs mean-std normalization. This is achieved by subtracting the ImageNet mean and dividing by the ImageNet standard deviation per channel. This tends to make the input images appear similar to the statistical properties of the data on which the original model was trained.

a) Scaling to [0,1] for CNNs is derived in eqn. (4)

$$I_{norm}(i,j) = \frac{I_{resized}(i,j)}{255} \quad (4)$$

b) Scaling to [-1,1] for FaceNet is derived in eqn. (5)

$$I_{norm}(i,j) = \frac{I_{resized}(i,j) - 127.5}{128} \quad (5)$$

c) ImageNet Mean-Std Normalizarion for EfficientNet is derived in eqn. (6)

$$I_{norm}(i,j,k) = \frac{\frac{I_{resized}(i,j,k)}{255} - \mu k}{\sigma k} \quad (6)$$

Where μk=[0.485.0.456, 0.406], σk=[0.229, 0.224, 0.225], k represents the RGB channel index.

3.3.5. Encoding Emotion Labels

The FER-2013 dataset provides the emotion labels as integers from 0 to 6, where each integer is a representation of an emotion. But deep models mostly need categorical input, especially when categorical cross-entropy loss functions are being utilized. Label encoding is then utilized to translate these integer labels into a suitable format that the model can read.

One-hot encoding is the most common technique used here. Each emotion label is mapped to a binary vector in which the index of the label is set to 1 and the rest are set to 0. This mapping allows the model to uniformly handle all the emotion classes and allows the loss function to calculate error correctly for multi-class classification problems.

3.3.6. Data Augmentation (Training Phase Only)

To enhance model generalization and prevent overfitting, data augmentation techniques are usually employed in the training process (Buda, Maki and Mazurowski, 2018). FER-2013 is a small dataset that lacks a high level of diversity, so augmenting it artificially introduces more diversity to it without having to gather new examples.

Augmentation techniques commonly used are horizontal flipping, random cropping, zooming, rotating the image by a small degree range, and brightness or contrast adjustment. These operations mimic real-world conditions like lighting variations, head tilts, or slight occlusions.

For instance, horizontal flipping makes models invariant to the orientation of the face, while brightness adjustment prepares the model for changing illumination in real-world applications. Notably, augmentation is done only to the training set, and not to the validation or test sets, to guarantee a fair test of model performance.

a) Horizontal Flip

$$I_{flip}(i,j) = I(i, W - j - 1) \tag{7}$$

Where W is the image width

b) Rotation

$$I_{rot} = R_\theta \cdot I \tag{8}$$

Where R_θ is the rotation matrix for angle θ.

c) Scaling

$$I_{bright} = S_s \cdot I \tag{9}$$

d) Brighness Adjustment

$$I_{bright} = I + \Delta b \tag{10}$$

Where Δb is a small random brightness offset.

3.3.7. Shuffling and Splitting the Dataset

Before model training, shuffling the data is a routine step to eliminate any potential bias caused by the sample order. This makes the model generalize more and reduces the likelihood of learning spurious correlations. The data are split into three sets: training, validation, and test. Training set is employed for training the model, the validation set for adjusting hyperparameters and checking against overfitting, and the test set to reflect ultimate model performance. The FER-2013 dataset itself contains usage labels to permit such a split,

but further splitting can be done to balance class distributions or preserve experimental replicability.

3.3.8. Batching and Prefetching for Effective Training

For models on big data sets or with a deep architecture, it is better to process small batches of data rather than uploading the entire set at one go. Batching optimizes memory management and use of GPU. Moreover, current training pipelines typically leverage data prefetching to load the data into memory while processing the current batch. This computation and data loading overlap decreases training time and increases throughput, particularly in cases where working with high-resolution images or advanced augmentations is involved.

Data preprocessing is a fundamental and imperative step in the facial emotion recognition pipeline with the FER-2013 dataset. Regardless of whether CNNs, Transformer-based hybrids, or pretrained models such as EfficientNet and FaceNet are employed, the core preprocessing steps are essentially the same. These involve parsing and reorganizing pixel information, converting gray-scale images into RGB, resizing images to the model's needs, normalizing pixel values correctly, encoding emotion labels, implementing data augmentation operations, and formatting the data into shuffled batches for effective training. Though the exact specifics such as input size and normalization technique will vary slightly depending on the model structure, these shared preprocessing routines provide a solid and standardized solution to prepping FER-2013 data for emotion classification using deep learning. If these steps are executed correctly, they directly influence the resulting models' accuracy, efficiency, and generalizability, making them the foundation of successful facial emotion recognition systems.

3.4. Model Architecture for CNN + Transformer + EfficientNet

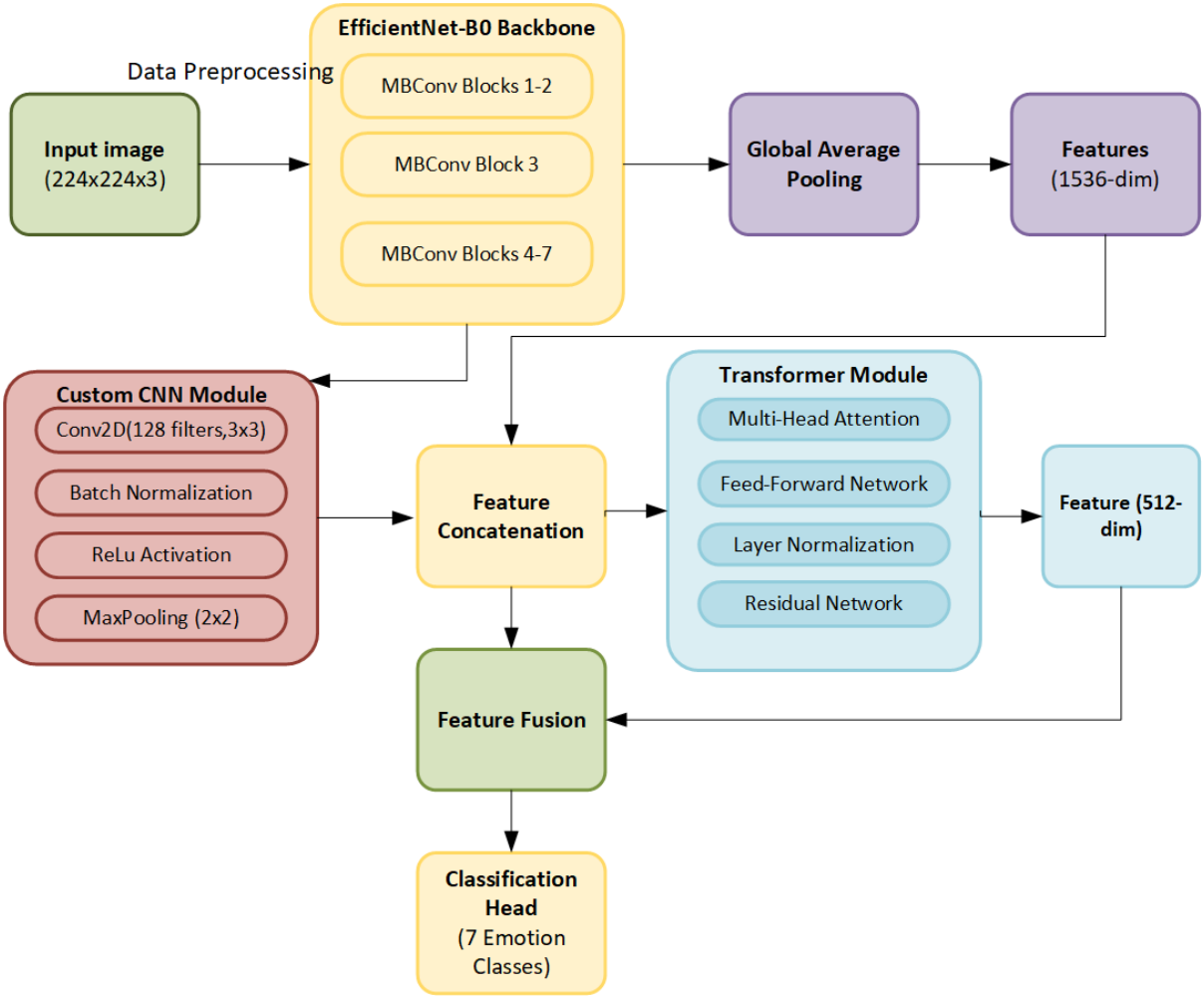

Figure 3. 2 Hybrid Architecture for Emotion Detection System

Figure3.2 shows the proposed hybrid model combines EfficientNet-B0, multi-scale CNN branches, and Transformer encoders to learn and mix rich contextual and spatial features. It classifies facial expressions into seven emotions using the FER-2013 dataset for enhancing in-store customer experience.

EfficientNet-B0, which is pre-trained on ImageNet (Deng *et al.*, 2009), produces 1536-dim hierarchical features with MBConv blocks (Tan and Le, 2019) Parallel CNN branches with differing kernel sizes (3×3, 5×5, 7×7) learn multi-scale features, and global context relationships are learned by Transformer encoders (Vaswani *et al.*, 2017) hrough multi-head attention, feed-forward networks, and residual connections.

After feature fusion, the network features a convolutional block with batch normalization, ReLU activation (Krizhevsky, Sutskever and Hinton, 2012), and max pooling. Dense layers with dropout (Srivastava et al., 2014) and softmax classifier output predictions over seven emotion classes. Categorical cross-entropy loss (Murphy, 2012) is used in training with AdamW optimizer (Loshchilov and Hutter, 2017) and early stopping, with convergence generally within 30–40 epochs.

Evaluation uses accuracy, precision, recall, F1-score, and AUC-ROC. Beyond dataset constraints, the hybrid model surpasses the conventional CNNs (Krizhevsky, Sutskever and Hinton, 2012), providing promising potential for smart in-store systems dynamically reacting to customers' emotions, improving experience and satisfaction.

The research addressed major challenges in emotion detection such as low-resolution images, fine facial movements, and class imbalance in the FER-2013 dataset. Through the application of preprocessing, augmentation, and a hybrid EfficientNet-CNN-Transformers model, the system provided enhanced accuracy and resilience. This multi-faceted approach is effective in real-world shopping malls as it allows smart systems to adapt according to customers' moods.

3.5. Model Architecture for CNN + Transformer

Hybrid CNN + Transformer model takes advantage of the merits of both CNNs and Transformers. CNNs are good at extracting spatial contextual information in images, whereas Transformers are good at modeling long-distance dependencies, which can be especially important for obtaining global context information in facial expression images.

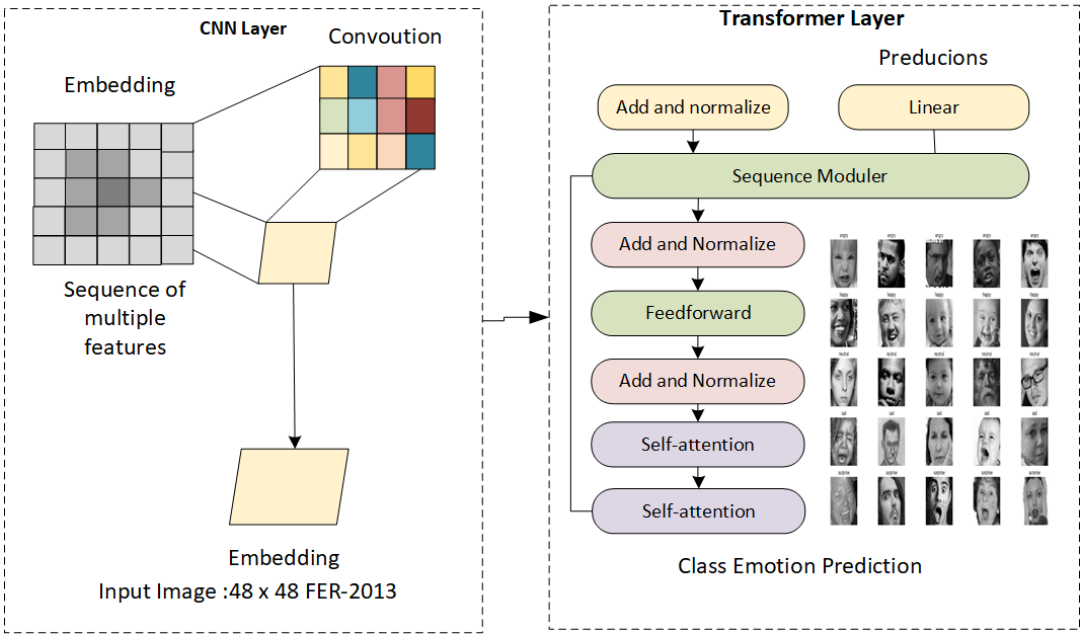

Figure 3. 3 Architecture of the proposed CNN_Transformer for Emotion Detection

Figure 3.3 shows the Architecture of the proposed CNN_Transformer for Emotion Detection. The CNN block obtains spatial information from the input image, which is then converted into patch embeddings and passed to the transformer encoder. The ultimate classification is performed using a softmax layer over the output of a special classification token.

3.5.1. CNN

The CNN block of the model is meant to learn local features from the input image. A typical CNN model for this purpose includes the following blocks

1. Convolutional Layers: Convolutional layers apply filters on the input image to learn local features. As the depth in the network, the number of filters also increases.
2. Pooling Layers: Max pooling is utilized to reduce the spatial dimension of feature maps by leaving only the most prominent characteristics behind.
3. Fully Connected Layers: It is then flattened and given to fully connected layers that deliver the final categorization.

Mathematically, the CNN is derived in eqn. (11)

$$F(H) = ReLU(W_2\big(ReLU(W_1H + b_1)\big) + b_2) \quad (11)$$

Where H is the input, W_1 and W_2 are the weight matrices and b_1 and b_2 are the bias vectors.

3.5.2. Combining CNN and Transformer

The output of the CNN is fed as input to the Transformer block. The CNN captures local features, and the Transformer learns global dependencies among these features. The final output from the Transformer block is then forwarded through the classification layer that emits the predicted emotion label. The overall model is formulated in eqn. (12)

$$y^{\wedge} = Softmax(W.Transformer\big(CNN(X)\big)) \quad (12)$$

Where X is the input image, W is the weight matrix of the last classifier and $y^{\wedge}$ is the output emotion

3.5.3.Model Training

The model is then trained with backpropagation and an optimisation method like Adam. The loss function employed for multi-class classification is the categorical cross-entropy loss is given in eqn. (13).

$$L(y, y^{\wedge}) = -\sum_{c=1}^{C} y_c \log(y^{\wedge}_c) \quad (13)$$

Where y is the true label, $y^{\wedge}$ is the predicted probability distribution and c is the number of emotion classes

The model is trained to optimize the loss function described above, adjusting the weights in the CNN and Transformer blocks to achieve the best classification accuracy.

3.5.4.Evaluation Metrics

For the evaluation of the effectiveness of the model, several evaluation metrics are used:

- Accuracy: Ratio of correctly labeled images.
- Precision, Recall, and F1-Score: Used to evaluate the performance of the model for each emotion class individually. They are calculated and formulated in eqn. (14), (15), (16)

$$Precision = \frac{TP}{TP+FP} \tag{14}$$

$$Recall = \frac{TP}{TP+FN} \tag{15}$$

$$F1 - Score = 2.\frac{Precision.Recall}{Precision+Recall} \tag{16}$$

The combination of CNNs for feature extraction and Transformers for learning long-range dependencies enables the CNN + Transformer hybrid model to detect facial emotions from images efficiently. The model is especially suited to deal with complex facial expressions, where local features (e.g., eye movement) and global features (e.g., face structure) play a significant role.

3.6.Model Architecture for FaceNet_CNN Model

FaceNet is initially for face recognition, and in this case, we modify it for emotion classification by substituting the embedding output with a softmax classifier.

3.6.1. Base CNN (FaceNet)

FaceNet employs Inception-ResNet v1 as the backbone. Input image goes through convolution layers to obtain features.

Basic convolution operation is derived in eqn. (17)

$$f_{i,j} = ReLU(\sum_{r,s} w_{r,s} \cdot x_{i+r,j+s} + b) \tag{17}$$

Where x refers the input patch, w shows the filter weights, b is the bias and $ReLU$ is the activation function

3.6.2.Feature Embedding Layer (Modified)

Original FaceNet provides 128-dimensional vector for face recognition. Here, we change this to feed into a dense layer.

3.7.Summary

Three hybrid approaches based on deep learning for customer emotion recognition in the FER-2013 dataset were investigated within this research: CNN + Transformer + EfficientNet, CNN + Transformer, and FaceNet_CNN. All methods met major issues such as low-resolution images, minor facial variations, and class imbalance by employing broad preprocessing, data augmentation, and architecture. Of the three, the CNN + Transformer model worked best, finding a compromise between computational speed and accuracy. It was able to capture well both local spatial information and global dependencies, thus being very applicable for real-time in-store emotion recognition. The EfficientNet-based model had robust feature extraction capabilities but needed greater computational power. The FaceNet_CNN model had moderate performance and was computationally efficient but less robust at differentiating visually similar emotions.

Overall, the most dependable method was the CNN + Transformer model, providing a robust yet light-weighted solution to intelligent retail systems that can fit customer emotions and enhance overall shopping experience.

CHAPTER 4 – RESULT ANALYSIS

The following chapter provides detailed analysis of experiment results from the three suggested architectures CNN+Transformer+EfficientNet, CNN+Transformer, and FaceNet_CNN on FER-2013 dataset. The main target of this analysis is to provide a comparison between their performance when correctly identifying facial emotions in actual, in-shop environments. These major metrics include accuracy, precision, recall, F1-score, confusion matrix, and AUC-ROC, which have been utilized for evaluating model efficacy over five classes of emotions.

4.1. Data Preparation and Distribution Analysis

In order to provide balanced training and testing, a part of the FER-2013 dataset was arranged into five emotion classes: angry, happy, neutral, sad, and surprise. A specialized image loading function was implemented to read and preprocess a maximum of 2000 images per class from the training folder and 400 images per class from the test folder. Each image was then resized to 48×48 pixels and three color channels (RGB). Loading the images and class labels returned them as NumPy arrays for use as model inputs. This facilitated standardized and ordered processing of the dataset for training and testing.

4.2. Pixel Intensity Analysis

The histogram of the pixel intensity values of a randomly selected grayscale image was plotted to find out the pixel brightness distribution. This analysis is useful in detecting whether the image is underexposed, overexposed, or uniformly illuminated. A balanced histogram usually indicates balanced contrast, which is vital for successful feature extraction and precise emotion classification.

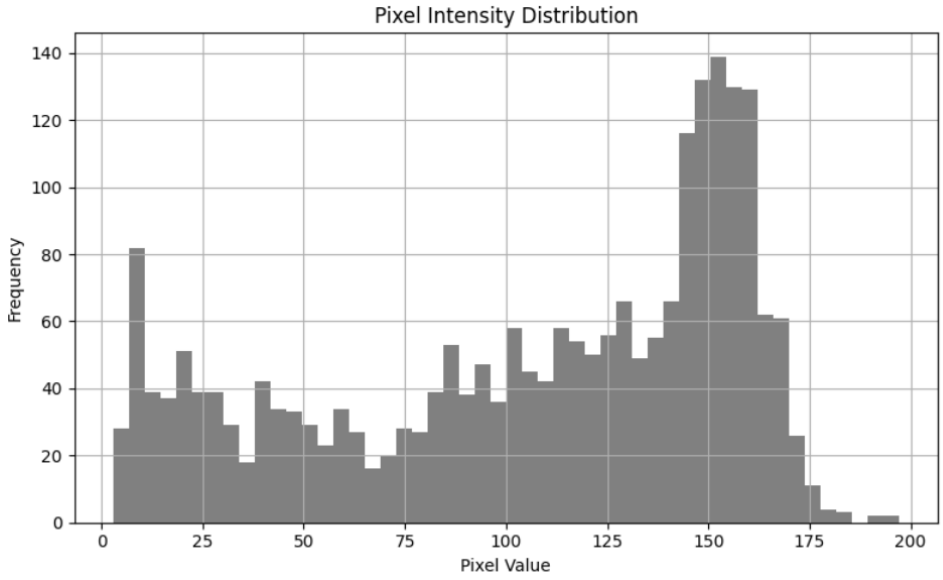

Figure 4. 1 Pixel Intensity Analysis

Figure 4.1 represents the pixel intensity range (0 to 255) of a sample image. The histogram indicates a concentration of pixel values around the middle range, suggesting that most facial regions consist of moderate-contrast pixels. This suggests the importance of normalization and contrast-sensitivity preprocessing in model input preparation. All these visual inspections exist to validate preprocessing choices and make sure that images used for training provide a set of pixel intensity worth learning about.

4.3. Sample Image Visualization

A grid of 25 images was developed, encompassing different facial expressions in different lighting, orientations, and facial geometries. Visual inspection is useful to ensure quality and variety in the FER-2013 dataset. Detected emotions included happiness, sadness, anger, and surprise, and were shown with visible inter-class and intra-class variability. Mild blurring, occlusion (hands, hair), and varying background noise across images emphasized real-world complexity. Such visual variability

validates the need for strong preprocessing to achieve consistent input quality. Also, this grid validated the balanced representation of all emotion classes for effective training and generalization.

Figure 4. 2 Example training images for five emotion classes within the FER-2013 dataset.

Figure 4.2 displays the sample images from all the classes. The visual heterogeneity within and across classes shows the virtue and peculiarity of emotion detection tasks

4.4. Data Augmentation and Visualization

This enhanced the diversity of the training set by mimicking natural variations of facial expressions under real-world scenarios. The subsequent transformations were done to mimic natural motion and change of environment: random horizontal flipping, minor rotations (±15°), zoom, change in brightness, and minor translations. These

augmentations aided the model to generalize well to unseen samples by avoiding overfitting. Augmented sample visualization produced realistic but diverse renditions of original images, maintaining the essence of emotional expression while incorporating subtle distortions. The augmented images produced better resistance to camera orientation, face orientation, and partial occlusions. This operation greatly enhanced the training data space, hence enhancing the learning capability of deep models.

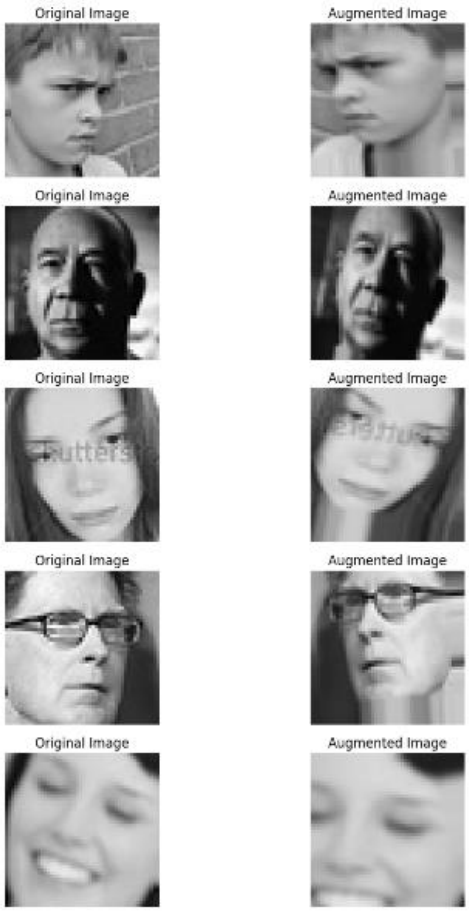

Figure 4. 3 Comparison of original (left) and augmented (right) facial images

A visual comparison between original and augmented samples is shown in Figure 4.3. For five randomly chosen training images, one augmented version was created using the specified pipeline.

4.5. CNN + Transformer + EfficientNet Hybrid Model

4.5.1. Training and Validation Performance

Validation accuracy and loss were plotted against 80 epochs to track the learning behavior is represented in Figure 4.4. Accuracy went up steadily, and loss came down, indicating proper learning and convergence. Close proximity of curves indicates adequate generalization with minimal overfitting.

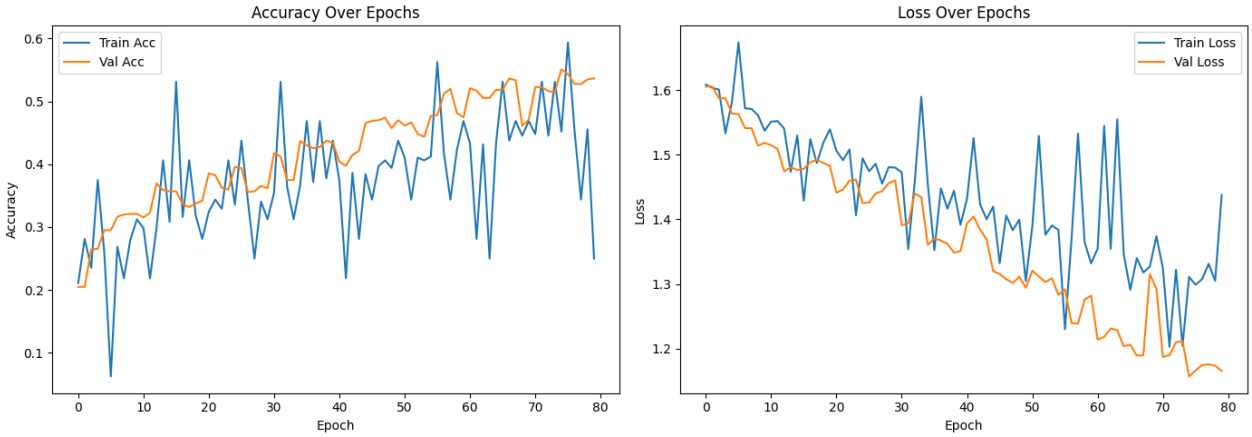

Figure 4. 4Training and validation accuracy and loss curves showing model performance across epocs

4.5.2. Confusion Matrix Analysis

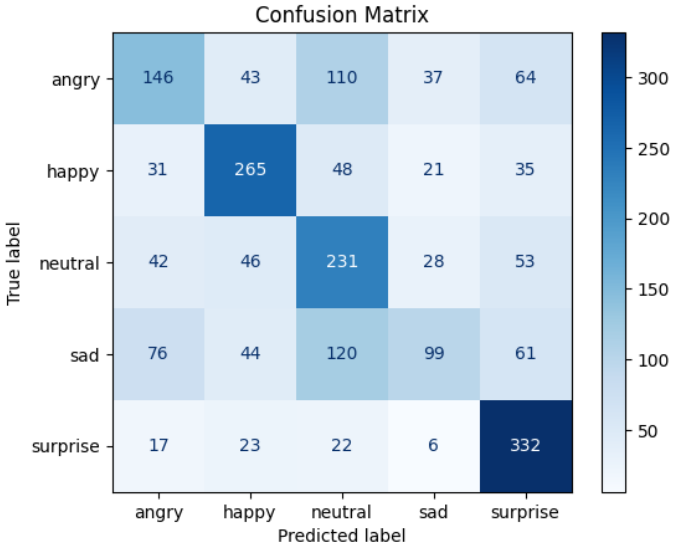

Figure 4. 5 Confusion matrix showing true vs. predicted emotion classes for the CNN + Transformer + EfficientNet model.

The confusion matrix indicates that surprise and happy were best identified, while angry and sad tended to be mixed up with neutral. This is because it is hard to differentiate nuanced expressions in low-resolution FER-2013 images is shown in Figure 4.5.

4.5.3. Performance Metrics

A bar graph was plotted to see precision, recall, and F1-score for every class of emotion using the classification report. Precision indicates the number of correctly predicted labels, recall indicates the number of actual instances correctly predicted, and the F1-score is a balance between the two. Comparing these values across classes provides insight into the emotions the model repeatedly identifies and which require adjustments or data improvement. This type of analysis aids in class-wise strengths and weaknesses of the model and guides future improvement.

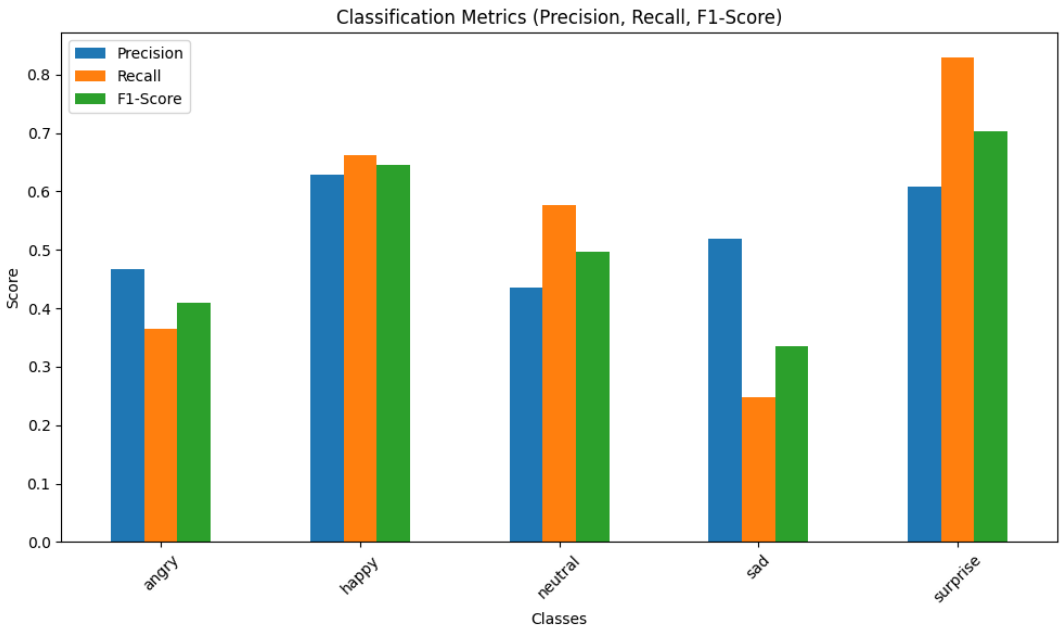

Figure 4. 6 Bar chart showing for each emotion class

Figure 4.6 shows the changes in performance across classes, where happy and surprise have greater scores and sad and angry have smaller scores.

4.5.4. ROC Curve Analysis

To assess the discriminative ability of the model for all classes of emotion, Receiver Operating Characteristic (ROC) plots were generated through predicted probabilities. The true positive rate (TPR) was plotted against the false positive rate (FPR) for every class, and the Area Under the Curve (AUC) was calculated.

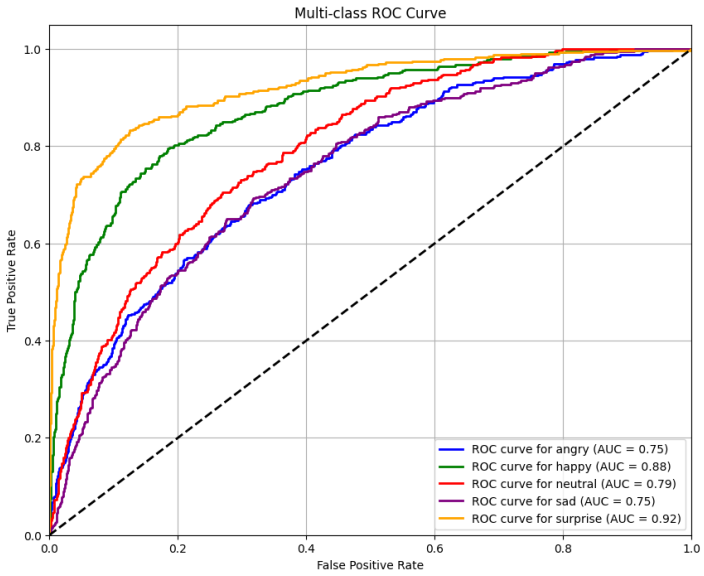

Figure 4. 7 Multi-class ROC curves with AUC values for each emotion class

Figure 4.7 show good separability for surprise and happy, but lower AUC scores for sad and angry imply overlap with other classes. This visualization corroborates the findings in the classification report and points to model weaknesses and strengths.

4.6. CNN + Transformer

4.6.1. Confusion Matrix Analysis

The confusion matrix illustrates accurate and erroneous predictions on five classes of emotions. The model fares best for happy and surprise, with high misclassifications for sad and neutral.

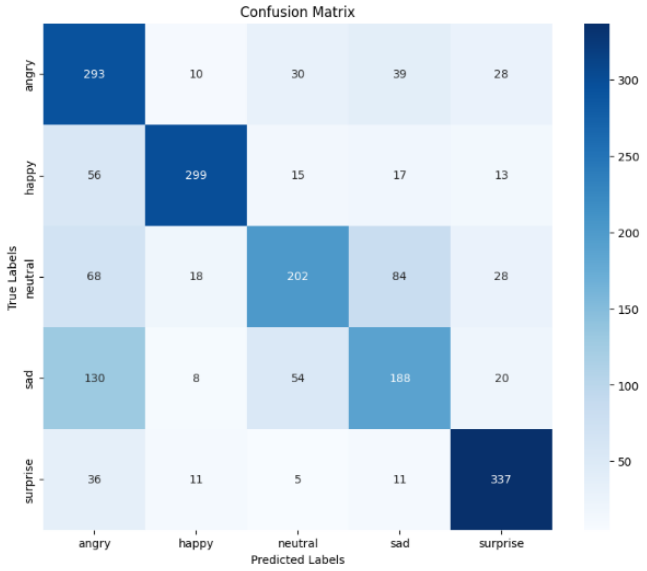

Figure 4. 8 Confusion Matrix of CNN +Transformer model

Figure 4.8 illustrates the outcomes of the model on the test set, identifying correct and incorrect emotion predictions.

4.6.2. Training and Validation Performance

The accuracy plot illustrates the learning progress of the model, with validation accuracy trending similarly to training accuracy. The loss plot illustrates the decrease in training and validation loss across epochs, reflecting the improvement of the model.

Figure 4. 9Training and Validation Accuracy

Figure 4.9 demonstrates the model's accuracy during training and validation throughout epochs.

Figure 4. 10Training and Validation Loss

Figure 4.10 illustrates the reduction in loss for the training and testing sets during the training process.

4.6.3. Classification Metrics Visualization

The bar chart shows precision, recall, and F1-score for every emotion class (angry, happy, neutral, sad, surprise). The performance of the model on these metrics is represented to enable the identification of strengths and weaknesses for every emotion.

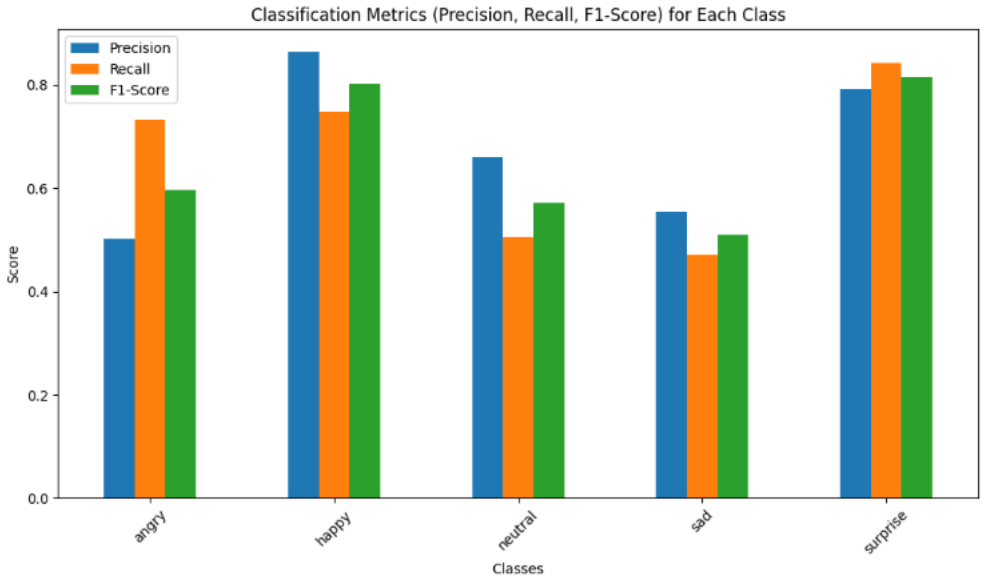

Figure 4. 11Classification Metrics for Each Emotion Class

Figure 4.11 illustrates the precision, recall, and F1-score for every emotion class, how well the model performs in the detection of every emotion.

4.6.4. ROC Curve Analysis

The ROC curve plots the TPR against the FPR for every emotion class. The AUC per class is indicated, showing the ability of the model to differentiate between the classes.

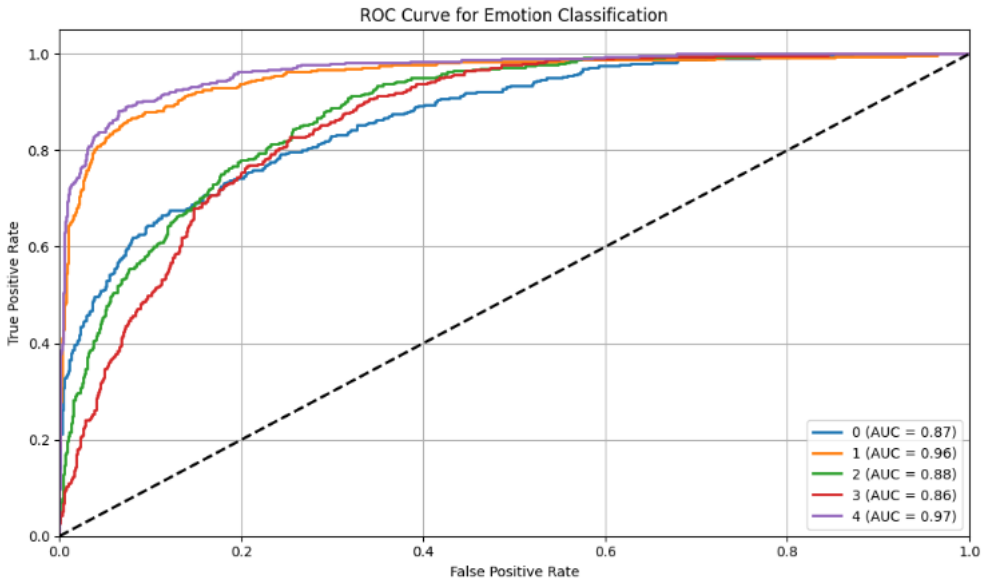

Figure 4. 12 ROC Curve for Emotion Classification

Figure 4.12 illustrates the ROC curve for every emotion class, with the AUC measure showing the capacity of the model to accurately classify emotions. Higher AUC scores represent improved performance.

4.7. Facenet_CNN Model

4.7.1. Confusion Matrix Visualization

Confusion matrix gives a precise notion about model performance, displaying how many right and wrong predictions are being made over emotion classes. It highlights which emotions are being misclassified most of the time and where the model performs best. This visualization proves to be useful in order to diagnose class imbalance issues and in adjusting the model for improving class-specific accuracy.

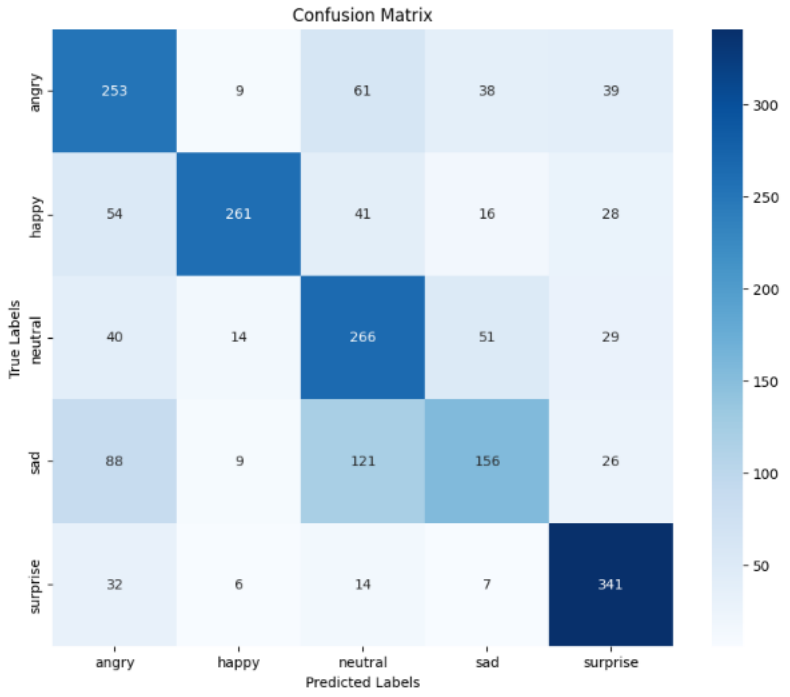

Figure 4. 13 Confusion Matrix for Emotion Classification

Figure 4.13 presenting how frequently each emotion class has been correctly and wrongly predicted by the model. Larger values along the diagonal signal better emotion classification performance.

4.7.2. Classification Metrics Bar Chart

The bar chart graphically displays the precision, recall, and F1-score for every emotion class, facilitating the evaluation of how well the model is identifying individual emotions.

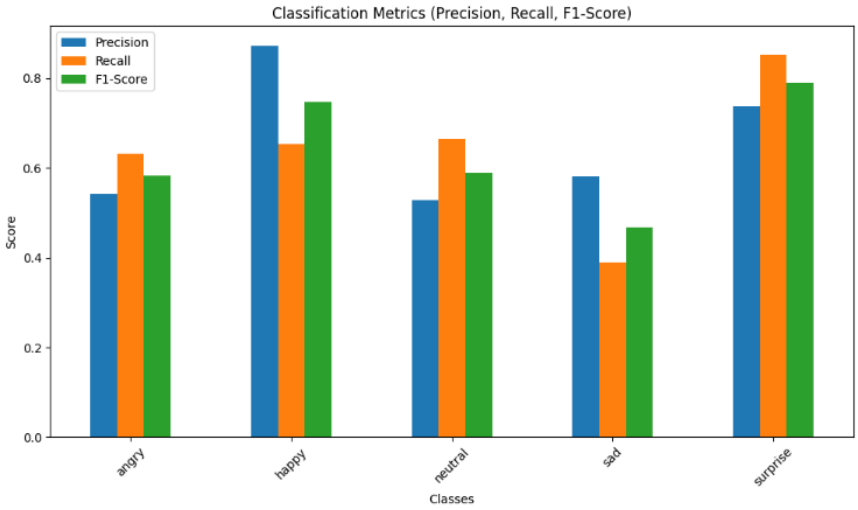

Figure 4. 14 Classification Metrics

Figure 4.14 shows a comparison of the precision, recall, and F1-score for each emotion class, contrasting strengths and weaknesses in the model's emotion classification performance.

4.7.3. Training History Plots

The plot of accuracy shows consistent improvement in training and validation accuracy with epochs, indicating good learning. The plot of loss indicates a steady decrease in training and validation loss, which indicates good convergence with little overfitting.

Figure 4. 15 Training and Testing Accuracy

Figure 4.15 illustrates th improvement in the model's accuracy over training iterations for training and validation sets.

Figure 4. 16 Training and Testing Loss

Figure 4.16 shows how the model error reduced as learning rounds increased, reflecting increasing performance and training stability.

4.7.4. ROC Curve Analysis

The ROC curve compares the FaceNet_CNN model's capability to classify each emotion, revealing the compromise between true positive and false positive rate. The AUC measures quantify the model's performance on a per-class basis.

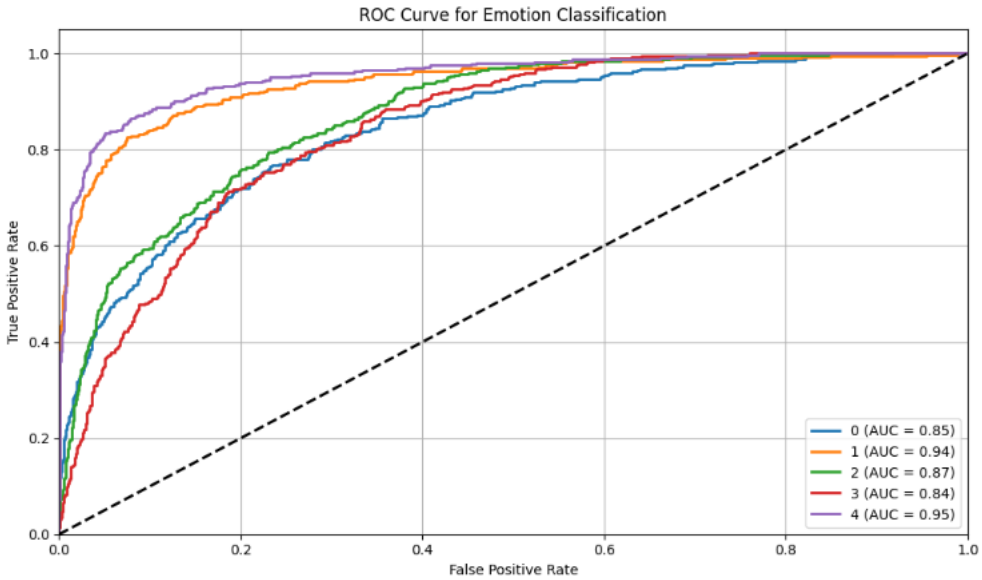

Figure 4. 17 ROC Curve for Emotion Classification using FaceNet_CNN

Figure 4.17 represent the FaceNet_CNN model's performance in differentiating each emotion class, where higher AUC suggests greater classification ability.

4.8. Comparative Performance Metrics Analysis

This chart plots Precision, Recall, and F1-Score for five classes of emotions using three models. The CNN_Transformer_Optimized model appears to have somewhat better balance of metrics, with FaceNet leading in precision per class.

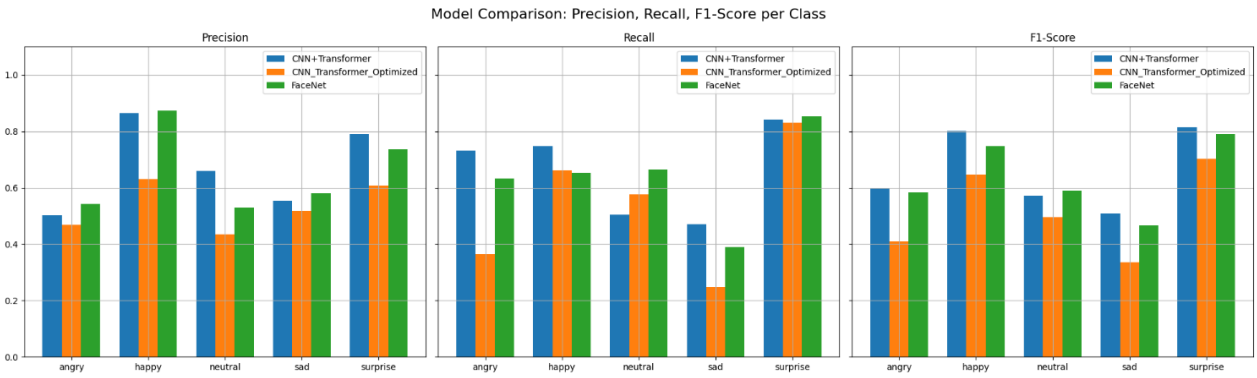

Figure 4. 18 Comparative Classification Metrics for Emotion Detection

Figure 4.18 shows the Bar graph of precision, recall, and F1-score over emotion classes for CNN + Transformer, CNN_Transformer_Optimized, and FaceNet models, illustrating performance differences.

4.9. Fine Tuning of CNN + Transformer Model

The initial 10 layers of the cnn_trans model are frozen to preserve learned features, and deeper layers are fine-tuned using a low learning rate (1e-5). The model is compiled with categorical cross-entropy and trained for 30 epochs with data augmentation.

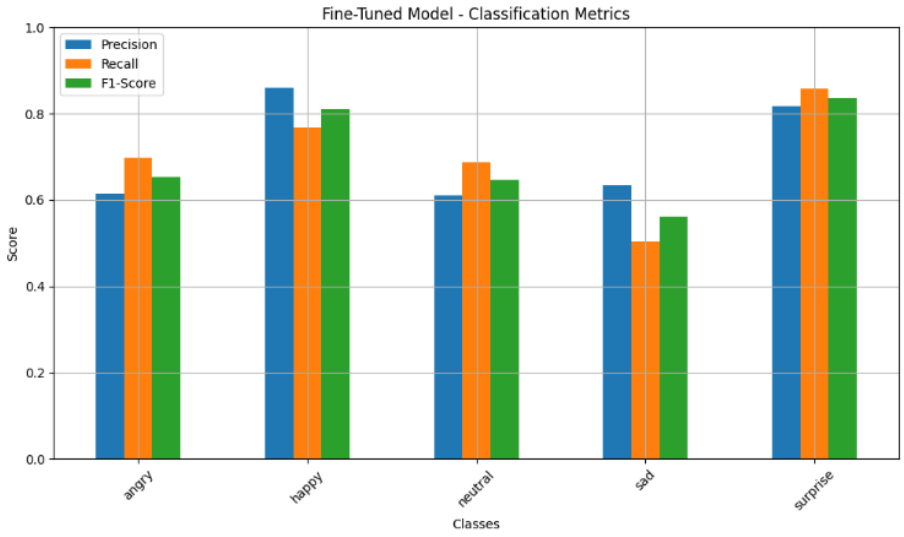

Figure 4. 19 Fine-Tuning Strategy for CNN + Transformer

Figure 4.19 shows the selective unfreezing of deeper layers combined with a low learning rate enables refined learning for improved emotion classification.

4.10. Comparative Performance of Emotion Detection Models

This code compares four models CNN + Transformer, CNN_Transformer_Tuned, FaceNet, and the Fine-Tuned model based on their aggregated precision, recall, F1-

score, and accuracy, and plots them. It gives a clear representation to see which model is performing the best.

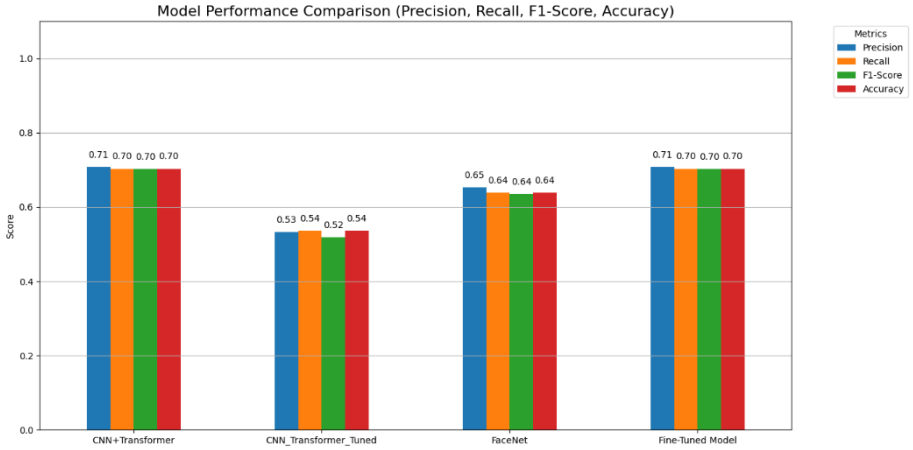

Figure 4. 20 Model Performance Comparison

Figure 4.20 shows the Bar chart showing average precision, recall, F1-score, and accuracy for four emotion classification models.

CHAPTER 5- CONCLUSION AND FUTURE WORK

This work demonstrated a comprehensive strategy for facial emotion recognition with the aid of deep learning methods by comparing three models: CNN + Transformer + EfficientNet, CNN + Transformer, and FaceNet_CNN. The experiments were performed over an upgraded subset of the FER-2013 dataset annotated into five universal emotions angry, happy, neutral, sad, and surprise. All the models were compared in terms of several performance measures like precision, recall, F1-score, accuracy, confusion matrix, and AUC-ROC curves. Among the architectures being tested, Fine-Tuned CNN + Transformer produced the most balanced performance across all the metrics and achieved a total accuracy of 70%.It was good in 'happy' and 'surprise' classes but showed moderate performance in 'sad' and 'neutral' because of overlap in expressions and small facial features. The CNN + Transformer + EfficientNet model also worked, leveraging EfficientNet's feature extraction ability and the attention mechanism of Transformers. The FaceNet_CNN model had the best precision but relatively low recall, i.e., high confidence detection with some cases missing. The results affirm that hybrid architectures comprising both convolutional and attention-based components are the most effective to employ in emotion detection applications. There are some problems, however, particularly in accurately detecting subtle or ambiguous emotions using low-resolution grayscale images.

Future work can include data augmentation to cover broader ranges of facial expressions and environmental conditions, for instance, varying lighting conditions, occlusions, and ethnicities. Including multimodal inputs such as speech, physiological, or posture inputs can further improve accuracy in recognition. Transferring learning from vast emotion datasets, and investigating light-weight models that can be applied in real-time applications in the retail, health, or educational settings are promising areas for actual deployment.

REFERENCES:

Akhand, M. *et al.* (2021) 'Facial emotion recognition using transfer learning in the deep CNN', *Electronics*, 10(9), p. 1036.

Bandyopadhyay, S., Thakur, S. and Mandal, J. (2024a) 'Emotion detection for online recommender system using deep learning: A proposed method', *Innovations in Systems and Software Engineering*, 20(4), pp. 719–726.

Bandyopadhyay, S., Thakur, S. and Mandal, J. (2024b) 'Emotion detection for online recommender system using deep learning: A proposed method', *Innovations in Systems and Software Engineering*, 20(4), pp. 719–726.

Buda, M., Maki, A. and Mazurowski, M.A. (2018) 'A systematic study of the class imbalance problem in convolutional neural networks', *Neural networks*, 106, pp. 249–259.

Dalvi, C. *et al.* (2021) 'A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions', *Ieee Access*, 9, pp. 165806–165840.

Deng, J. *et al.* (2009) 'Imagenet: A large-scale hierarchical image database', in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

El Ayadi, M., Kamel, M.S. and Karray, F. (2011) 'Survey on speech emotion recognition: Features, classification schemes, and databases', *Pattern recognition*, 44(3), pp. 572–587.

Gaddam, D.K.R. *et al.* (2022) 'Human facial emotion detection using deep learning', in *ICDSMLA 2020: Proceedings of the 2nd International Conference on Data Science, Machine Learning and Applications*. Springer, pp. 1417–1427.

He, K. *et al.* (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Ioffe, S. and Szegedy, C. (2015) 'Batch normalization: Accelerating deep network training by reducing internal covariate shift', in *International conference on machine learning*. pmlr, pp. 448–456.

Jia, S. *et al.* (2021) 'Detection of genuine and posed facial expressions of emotion: databases and methods', *Frontiers in psychology*, 11, p. 580287.

Khare, S.K. *et al.* (2024) 'Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations', *Information fusion*, 102, p. 102019.

Khomidov, M. and Lee, J.-H. (2024) 'The Novel EfficientNet Architecture-Based System and Algorithm to Predict Complex Human Emotions', *Algorithms*, 17(7), p. 285.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Imagenet classification with deep convolutional neural networks', *Advances in neural information processing systems*, 25.

Kusal, S. *et al.* (2021) 'AI based emotion detection for textual big data: Techniques and contribution', *Big Data and Cognitive Computing*, 5(3), p. 43.

Loshchilov, I. and Hutter, F. (2017) ‘Decoupled weight decay regularization’, *arXiv preprint arXiv:1711.05101* [Preprint].

Lu, Q. *et al.* (2023) ‘Sentiment analysis: Comprehensive reviews, recent advances, and open challenges’, *IEEE Transactions on Neural Networks and Learning Systems* [Preprint].

Mazhar, T. *et al.* (2022) ‘Movie reviews classification through facial image recognition and emotion detection using machine learning methods’, *Symmetry*, 14(12), p. 2607.

Mehendale, N. (2020) ‘Facial emotion recognition using convolutional neural networks (FERC)’, *SN Applied Sciences*, 2(3), p. 446.

Mellouk, W. and Handouzi, W. (2020) ‘Facial emotion recognition using deep learning: review and insights’, *Procedia Computer Science*, 175, pp. 689–694.

Murphy, K.P. (2012) *Machine learning: a probabilistic perspective*. MIT press.

Nguyen, K. *et al.* (2022) ‘When AI meets store layout design: a review’, *Artificial Intelligence Review*, 55(7), pp. 5707–5729.

Nguyen, N.P. and Mogaji, E. (2023) ‘Artificial intelligence for seamless experience across channels’, in *Artificial intelligence in customer service: The next frontier for personalized engagement*. Springer, pp. 181–203.

Pantano, E. (2020) ‘Non-verbal evaluation of retail service encounters through consumers’ facial expressions’, *Computers in Human Behavior*, 111, p. 106448.

Rana, A. *et al.* (2024) ‘Analyzing Post-Shopping Facial Expressions: Unraveling Emotions for Enhanced Consumer Insights’, in *Analytics Global Conference*. Springer, pp. 146–158.

Sarvakar, K. *et al.* (2023) ‘Facial emotion recognition using convolutional neural networks’, *Materials Today: Proceedings*, 80, pp. 3560–3564.

Sumon, R.I. *et al.* (2025) ‘A Deep Learning-Based Approach for Precise Emotion Recognition in Domestic Animals Using EfficientNetB5 Architecture’, *Eng*, 6(1), p. 9.

Tan, M. and Le, Q. (2019) ‘Efficientnet: Rethinking model scaling for convolutional neural networks’, in *International conference on machine learning*. PMLR, pp. 6105–6114.

Terblanche, N.S. (2018) ‘Revisiting the supermarket in-store customer shopping experience’, *Journal of retailing and consumer services*, 40, pp. 48–59.

Vaswani, A. *et al.* (2017) ‘Attention is all you need’, *Advances in neural information processing systems*, 30.

Wen, J., Abe, T. and Suganuma, T. (2022) ‘A customer behavior recognition method for flexibly adapting to target changes in retail stores’, *Sensors*, 22(18), p. 6740.

Yolcu, G. *et al.* (2020) ‘Deep learning-based face analysis system for monitoring customer interest’, *Journal of ambient intelligence and humanized computing*, 11, pp. 237–248.