BRAIN. Broad Research in Artificial Intelligence and Neuroscience
e-ISSN: 2067-3957 | p-ISSN: 2068-0473
Covered in: Web of Science (ESCI); EBSCO; JERIH PLUS (hkdir.no); IndexCopernicus; Google Scholar; SHERPA/RoMEO; ArticleReach Direct; WorldCat; CrossRef; Peeref; Bridge of Knowledge (mostwiedzy.pl); abcdindex.com; Editage; Ingenta Connect Publication; OALib; scite.ai; Scholar9; Scientific and Technical Information Portal; FID Move; ADVANCED SCIENCES INDEX (European Science Evaluation Centre, neredataltics.org); ivySCI; exaly.com; Journal Selector Tool (letpub.com); Citefactor.org; fatcat!; ZDB catalogue; Catalogue SUDOC (abes.fr); OpenAlex; Wikidata; The ISSN Portal; Socolar; KVK-Volltitel (kit.edu)
2026, Volume 17, Issue 3, pages: 77-90
Submitted: May 28th, 2026 | *Accepted for publication:* July 21st, 2026

Hybrid CNN-LSTM Framework-based Facial Emotion Recognition for Retail Decision Analytics

Narmadha R
Division of Commerce and International Trade,
Karunya Institute of Technology and Sciences,
Coimbatore, India.
narmadhar25@karunya.edu.in

Leo A*
Division of Commerce and International Trade,
Karunya Institute of Technology and Sciences,
Coimbatore, India.
leoa@karunya.edu

Saji Abraham
Karunya School of Management,
Karunya Institute of Technology and Sciences,
Coimbatore, India.
sajiabraham@karunya.edu
https://orcid.org/0009-0006-0726-2507

Shalini Divya Prasanna A
Division of Digital Sciences,
Karunya Institute of Technology and Sciences,
Coimbatore, India.
shalinidivya@karunya.edu

Janani T
Division of Commerce and International Trade,
Karunya Institute of Technology and Sciences,
Coimbatore, India.
jananit@karunya.edu.in

Abstract: *This facial emotion recognition (FER) framework, which is the basic system that integrates technologies and processes related to facial emotion recognition, will be used in a relevant research project—we will explore whether facial emotion recognition, the technology that judges current emotions based on facial expressions, can become a component of retail decision analysis tools that perceive emotions. If customers' emotions can be accurately identified, future retail decision support applications will gain an additional piece of reference information reflecting customers' behaviours. Most of the previous CNN-based FER systems focus on static frames for classification without considering the temporal evolution of emotions. In this paper, three deep learning models—CNN, Mini-Xception, and CNN–LSTM are compared in the context of facial emotion recognition for retail decision analytics. CNN and Mini-Xception are spatial methods, while CNN–LSTM combines spatial and temporal information. CNN models are pre-trained on the FER-2013 database (35,887 images) and CNN-LSTM is tested on the CK+ dataset (981 sequences). In the experimental environment of CK+, the CNN-LSTM model processes temporally ordered expression sequences and achieves an accuracy rate of 92.42%. This result proves that under this experimental setup of CK+, the method of modelling sequential time-series data is effective. As for the application of this framework to support retail decision analysis, it can only be regarded as a potential expansion direction at present—all relevant research on it is completed based on existing, reference benchmark facial expression datasets, and has never used transaction-related data generated in real retail scenarios, so it cannot be implemented in actual retail business for the time being.*

Keywords: *facial emotion recognition; CNN–LSTM; temporal attention; retail analytics; affective computing.*

How to cite: Narmadha, R., Leo, A., Abraham, S., Prasanna, S. D. A., Janani, T., Joseph, K. J., & Stepy, L. P. (2026). Hybrid CNN-LSTM framework-based facial emotion recognition for retail decision analytics. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience, 17*(3), 77-90. https://doi.org/10.70594/brain/17.3/5

Kevin Joseph J
Dept of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India.
kevinjoseph@karunya.edu.in,
https://orcid.org/0009-0009-6647-9839

Lourdu Stepy P
Karunya School of Management,
Karunya Institute of Technology and Sciences, Coimbatore, India.
lourdustepy@karunya.edu.in

* *Corresponding author: leoa@karunya.edu*

1. Introduction

Today, AI and computer vision, in conjunction with data analytics, are being used in various ways in the field of retail, including customer-experience and decision-support systems. The behaviour of the customers in such environments is important in understanding how to improve customer satisfaction, the layout of the stores and making effective business decisions. Human emotions are very important in determining shopping behaviour, purchase intent and the shopping experience. In general, conventional retail analytics tools mainly depend on transactional and customer feedback data, which covers responses in real time. Current facial emotion recognition advances have shown the possibility of deep learning models to determine emotional states based on facial expressions. However, most of the available methods are primarily focused on the recognition of static emotions and fail to consider the temporal dynamics of emotions in a dynamic shopping situation. A hybrid deep-learning system combining Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) is presented and used to recognise facial emotions and to model their temporal evolution in this paper. With the help of the FER-2013 dataset in pre-training and the CK+ dataset in validation of modelling of temporal emotional differences, facial expression analysis could provide another behavioural cue for next-generation emotion-aware retail analytics using the framework as a foundation to be explored in the next research work on the shopping decision-making processes in the intelligent retail environment. The present study is based on facial emotion recognition, which is not directly related to the real-world retail behaviours including purchase or sale, or dwell time, etc.

Facial Emotion Recognition (FER) is one of the main research areas in the affective computing domain, as it can recognise and process human emotions based on non-verbal facial expressions. Facial expression is one of the most spontaneous and natural ways of expressing emotion and is an important part of human interaction, perception, and behavioural response. Psychological and behavioural approaches to emotion, such as determining a limited set of universal expressions of emotion (happiness, anger, fear, sadness, surprise, disgust, neutrality), were a great source of inspiration for original research in FER. The findings of these studies were used as the theoretical foundation of the interpretation of facial motions as the expression of emotional conditions and their relationship to decision-making. However, the first operationalisation attempts were largely rule-based and relied heavily on manual feature extractors. These systems were rigid and very susceptible to changes in light, face and differences between people. This meant that these initial designs were not well scaled and not applicable to complex real-world settings, and their applicability and strength was limited in a real sense (Sariyanidi et al., 2015).

Common features used in typical FER methods include local features such as Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG) and Gabor features, as well as classifiers such as Support Vector Machines (SVM), k-Nearest Neighbours (KNN) and Naive Bayes. In the controlled environment, these methods can perform well but require manual labor for making these manually engineered features, making them less robust to illumination variations, pose, occlusion, and image quality. The features and textures of faces which are coded by the classifiers include Local Binary Pattern (LBP), Histogram of Oriented Gradient (HOG), and Gabor filters. These approaches had some moderate success with controlled data with little variation, but failed with uncontrolled data (Mellouk & Handouzi, 2020). Variations in illumination, facial occlusions, variation in head pose, and image noise to a great extent influenced classification accuracy. Moreover, these models were highly domain-specific and not easily used in new data sets or application scenarios due to the need for a massive amount of manual feature engineering. This led them to be poor in generalisation and scalability and thus to seek more powerful ways to learn (Mollahosseini et al., 2016).

The face emotion recognition field was greatly enriched by deep learning, especially in the area of automatic hierarchical learning of information from images. The automatic hierarchical learning of features directly from raw pixel data using Convolutional Neural Networks (CNNs) completely changed image-based classification. In contrast to conventional methods CNNs learn to

represent information discriminatively across the spatial dimensions without the need for any manually designed facial features, using convolutional, activation and pooling operations (Lopes et al., 2017). VGGNet, ResNet and Xception, which are popular CNN architectures, have been effectively trained on standard FER datasets, like FER-2013, CK + and JAFFE. These architectures can learn the intricate shapes of faces and the finer details of variations of expression, which are hard to express in a model. Consequently, CNN-based FER systems have always exhibited high accuracy, robustness, and flexibility in comparison with traditional machine learning methods (Hochreiter & Schmidhuber, 1997). FER-2013 is one of the most widely used datasets to test facial emotion recognition models based on deep learning in comparison with other benchmark datasets. The data are low-resolution grayscale facial images captured in the real world and there are some challenges such as noise, occlusion and lighting variation. However, there are some studies in the literature that show that CNN models trained from FER-2013 have good generalisation performance (Yu et al., 2018).

Both lightweight architectures like Mini-Xception have attracted special interest because of their ability to balance between efficiency and accuracy. The models exhibit an extremely low computational complexity with the low computation depth-wise separable convolutions and residual connections, as well as competitive recognition performance. There are some peculiarities that make these appropriate for real-time applications and resource-constrained environments (Rouast et al., 2021). There are many existing FER systems that are based on a static classification problem, and CNNs could be very efficient in extracting spatial features of the face. The frames of emotions are calculated independently of each other, in no particular order. While CNN-based FER models are able to model spatial facial features, the image-level classification does not make use of a temporal evolution of expressions explicitly. This limitation is particularly important with moving images, in which the gestures are moving from frame to frame. Using LSTM to complement the CNN-based spatial model, temporal models can then be used to learn the temporal relationships in facial expressions (Li et al., 2017).

To overcome the disadvantages of the above-mentioned approaches, scholars have been exploring the temporal modelling techniques using Recurrent Neural Network (RNN) framework, particularly Long Short-Term Memory (LSTM) networks (Li & Deng, 2019). The LSTM networks are the networks modelled with the data flow that is sequential in nature. These models have also shown to be more effective in detecting subtle, fleeting expressions of emotion which are difficult to detect by frame-based models (Fu et al., 2020). Networks with a hybrid CNN-LSTM architecture are an important improvement in affective computing because they unify both the spatial and the temporal learning within the same network. The CNNs' functions in these architectures are to extract high-level facial features of specific frames in the image, and to temporally model sequence of frames using LSTM layers (Li & Deng, 2022). These will enable the system to capture both the actual expression and how the expression changes. CNN-LSTM models have been found to be better than CNN-only models in some studies for extracting the emotion from video clips, analysis of gestures, and human actions. Embracing of temporal modelling also adds a high degree of robustness to systems operating in real-time and dynamic conditions in which the emotional patterns are constantly changed (Lucey et al., 2010).

Smart systems which react to emotions have been increasingly applied in a wide area, including health care performance, driving behaviour in cars, human computer interface and smart surveillance. In medical use, emotion recognition has been used in monitoring mental conditions of patients, and also in monitoring psychological stress. In the case of a car system, it can be used to detect driver fatigue and stress (Langner et al., 2010). The applications demonstrate the versatility and value of emotion recognition technology. Most of the current implementations are however aimed at emotion detection and classification as opposed to predictive analysis. Moreover, not much work has been conducted to investigate the implementation of emotion-sensitive systems in business or retail environments, although they can have an influence (Mohammadpour et al., 2017). Traditional methods of analysing shopping behaviour used in retail environments are based on

transactional records, customer survey, loyalty, and sales analytics. These methods can provide informative information about the post-purchase behaviour but are unable to capture the actual reactions of emotions, which are highly influencing the decision to purchase products in real time (Boz & Kose, 2018). Customers' emotions affect their buying decision, their stay time at the store, and their shopping experiences (Deshmukh & Patil, 2026). Traditional retail analytics rely on transactional data and are based on retrospective surveys, and are unable to reflect real-time affective reactions preceding buying actions (Zarezadeh & Rezaeian, 2016). Non-intrusive computer vision can be a potential tool for capturing such responses, as a modality called facial emotion recognition (Velammal et al., 2025).

2. Related Work

2.1. Customer Retail Analytics

Customer retail analytics is the systematic application of data and technology to analyse customer behaviour, preferences, emotional responses, and purchasing patterns in the retail space. Traditional methods rely mostly on transaction records, feedback from consumers, loyalty data and sales information. However, these approaches may not be able to capture real-time emotional responses of the customers. Retailers may utilise deep learning based Facial Emotion Recognition (FER) to analyse emotional states via facial expressions. The CNN models extract the facial spatial features, and Mini-Xception provides computationally efficient emotion recognition for real-time applications. CNN-LSTM also captures temporal information by analysing the changes in emotion in consecutive frames. Thus, the three models provide complementary perspectives for understanding customer emotions and supporting intelligent retail decision analytics.

2.2. Static Facial Emotion Recognition

In early FER systems, hand-crafted features such as LBP, HOG, Gabor filters were used, and supervised classifiers (SVM, KNN, Naive Bayes) were applied on the top of these features, using the Facial Action Coding System (FACS) descriptors as rules. These methods yielded a moderate level of accuracy under controlled lighting situations, but failed to perform well when dark, with pose changes and with occlusion. CNNs learned hierarchical spatial representations directly from pixel data in place of manual feature engineering, which eliminated the need to manually craft features for CNNs. CNNs were the most dominant method for static FER, with some pre-trained architectures like VGGNet, ResNet and Mini-Xception performing well on benchmark datasets (FER-2013, CK+, JAFFE). All CNN based FER systems, however, consider emotion recognition as a one-time classification problem and the information obtained from individual frames is processed independently. This framing eliminates temporal cues from emotional expression sequences, especially important for dynamic environments where the emotions continuously change.

2.3. Temporal Emotion Modelling

Sequential dependencies are captured in Long Short-Term Memory (LSTM) networks, which involve gated memory cells that can remember and forget information across time steps. Applied to video-based FER, the LSTMs are better at catching emotional changes between subsequent frames than in the classification of posed expression sequences frame by frame. Hybrid CNN–LSTM architectures integrate CNN to extract the spatial features and LSTM to model the temporal features within a unified framework. Temporal models are also enhanced by attention mechanisms that amplify emotionally relevant frames, while de-amplifying transitional frames. In several video-based facial-expression recognition tasks, spatial feature extraction and sequential modelling approaches, which combine CNN and LSTM networks, have been proven helpful in the modelling of the temporal component of the facial expression. Combining spatiotemporal feature extraction and sequential modelling, the hybrid CNN-LSTM approach has been proven to be effective for some video-based facial-expression recognition tasks.

3. Inferences from Literature Survey

A literature search was carried out to pinpoint some larger research gaps for the proposed application. The present study mainly tackles the first gap regarding the modelling of temporal emotions, whilst the other gaps are identified as avenues for future research: (1) The dynamics of emotion over time are not considered in retail FER applications; (2) No measurements have been made of the distribution shift between training data in the lab and deployment data in the retail environment; (3) There is no privacy preserving architectures that complies with existing data protection laws; (4) The causal validity of emotion-purchase inference has not been tested; (5) The risk of proxy corruption under sustained optimisation has not been analysed.

4. Proposed Methodology

4.1. System Overview

The framework comprises five sequential modules: (1) facial image acquisition, (2) preprocessing, (3) CNN or Mini-Xception spatial feature extraction, (4) CNN–LSTM temporal modelling, and (5) emotion-driven decision analysis. Fig. 1 illustrates the complete pipeline.

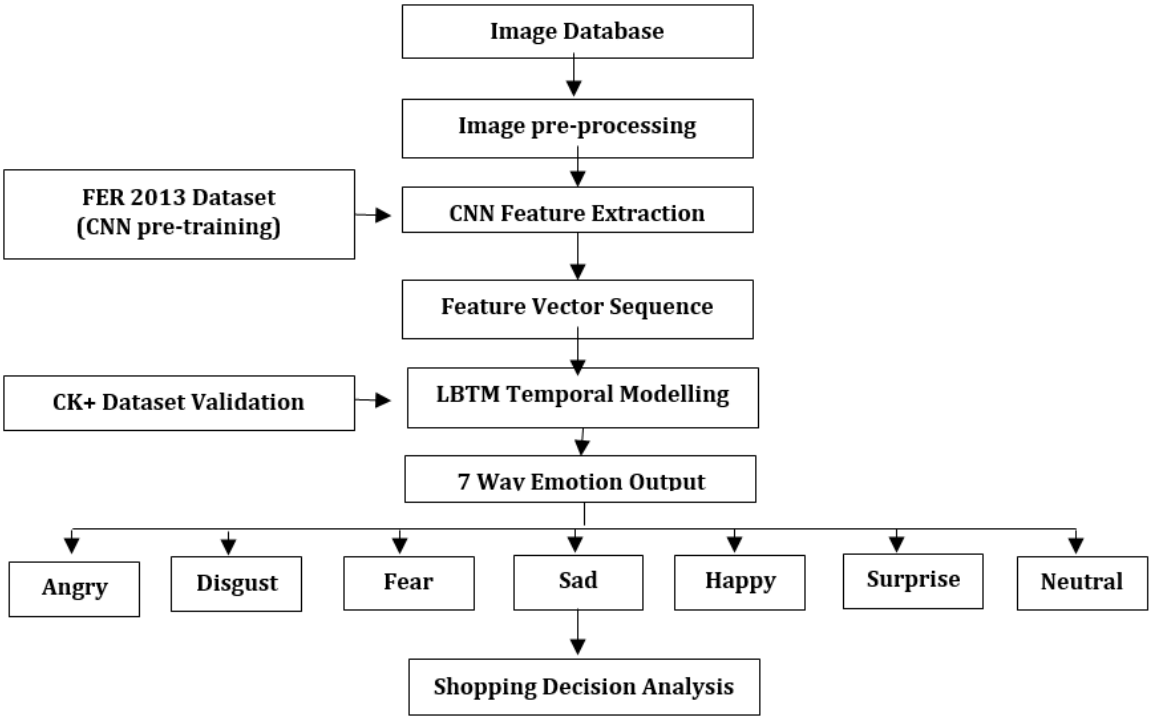

Figure 1. Proposed CNN–LSTM framework for facial emotion recognition and retail analytics

The proposed methodology presents a suggested approach that outlines a CNN–LSTM-based facial emotion recognition system intended to be a part of a broader retail decision analytics system that is emotion-aware. The process begins with the collection of input images. An image pre-processing stage to enhance data quality and consistency is used in this work. This stage includes operations such as image resizing, grayscale conversion, normalisation, and noise reduction. Pre-processing helps reduce the impact of real-world challenges such as illumination variations, background interference, and image distortions that are commonly encountered in shopping mall environments.

The pre-processed facial images are then passed to the CNN feature extraction module. In this stage, deep spatial facial features are extracted using a Convolutional Neural Network pre-trained on the FER-2013 dataset. Pre-training on FER-2013 allows the CNN to learn generic

and discriminative facial emotion patterns. The CNN focuses on capturing spatial characteristics such as eye movement, facial contours, and mouth expressions that are essential for emotion recognition. The extracted CNN features are organised into a feature vector sequence, which serves as input to the LSTM temporal modelling module. The LSTM network models temporal dependencies and emotional variations over time, allowing the system to track changes in a customer's emotional state as they interact with the shopping environment. The temporal modelling capability enables a transition from static emotion recognition to emotion prediction, which is critical for understanding customer behaviour in dynamic retail scenarios. The LSTM module is validated using the CK+ dataset, which contains well-annotated emotion sequences suitable for learning temporal emotion dynamics. After temporal modelling, the output of the LSTM network is passed through a SoftMax emotion classification layer, which assigns probabilities to each of the seven basic emotion classes. The system categorises facial expressions into angry, disgust, fear, sad, happy, surprise, and neutral, which are widely used in facial emotion recognition research and are consistent with the FER-2013 dataset. These emotion classes collectively provide a comprehensive representation of customer emotional states. The 7 emotion outputs for the previous 7 days may be analysed as inputs for exploration of potential relationships between the emotional trends and shopping behaviours in future. Sustained positive emotions such as happiness may indicate interest and purchase intent, while negative emotional patterns such as sadness, fear, or disgust may suggest disengagement or dissatisfaction. By analysing these emotional responses over time, the proposed system supports intelligent retail analytics, enhances customer experience understanding, and enables informed decision-making in smart shopping mall environments.

4.2. Image Datasets

The datasets used in this work are FER-2013 and CK+ (Extended Cohn–Kanade). The FER-2013 includes 35,887 greyscale images (48 × 48 pixels) across seven emotion categories (angry, disgust, fear, happy, sad, surprise, neutral), collected from Internet sources under uncontrolled conditions. Training set: 28,709 images. Validation set: 3,589 images. Test set: 3,589 images. The CK+ (Extended Cohn–Kanade) includes 981 image sequences from 123 subjects, with expressions progressing from neutral to peak intensity. The sequences are annotated with FACS Action Unit labels and emotion categories. For CK+, data were divided at the subject level prior to the generation of the temporal sequence to ensure that the training and test sets are subject independent. No subject added frames or temporal sequences to more than one partition.

The two data sets are not benchmarks per se, but are used to complement experiments. For spatial feature learning in an uncontrolled setting, FER-2013 offers a relatively large set of facial images, while for temporal modelling, CK+ offers a set of facial-expression sequences that are annotated. As such, the accuracy values for the two datasets are reported in the context of their experimental environment and are not necessarily seen as a direct comparison.

4.3. Preprocessing

Preprocessing standardises input data across datasets and deployment conditions: (1) face detection and cropping using Haar Cascade classifier; (2) greyscale conversion; (3) resizing to 48 × 48 pixels; (4) pixel intensity normalisation to [0, 1]; (5) Gaussian noise reduction (3 × 3 kernel); and (6) temporal sequencing into 10-frame windows for CNN–LSTM input (~0.3–0.5 seconds at 25–30 fps). The temporal sequence is modelled with 10 consecutive facial frames for each temporal sequence and fed into the CNN–LSTM model. Fig. 2 illustrates the preprocessing pipeline.

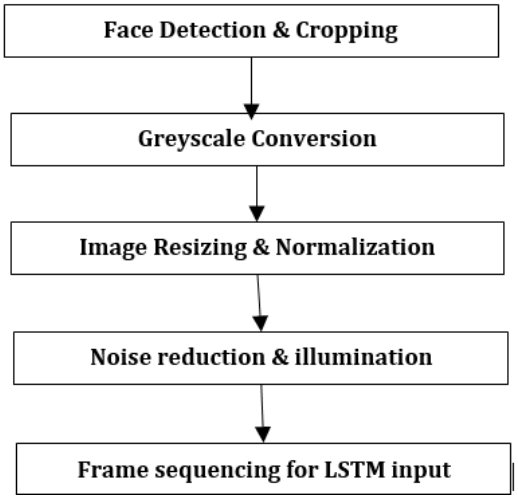

Figure 2. Preprocessing pipeline: raw frames undergo face detection, greyscale conversion, resizing (48 × 48), normalisation, noise reduction, and frame sequencing before model input.

Pre-processing is very vital to enhance accuracy and reliability of facial emotion recognition systems particularly in dynamic conditions like retail environments. Face detection is the initial process in the pre-processing pipeline that involves entailing facial areas and cutting them out of the input images or video frames. This step eliminates all the unnecessary background information and only the region of interest is sent to undergo further analysis. After the face has been recognised, the extracted face images are converted into grayscale format. Grayscale conversion simplifies computational tasks, and it provides compatibility with the FER-2013 dataset which also consists of grayscale pictures. After this, the resizing of all the faces into 48 x 48 pixels is done so that the dimensions of the hair, eyes, nose, and mouth of the faces remain the same to the CNN model. The image size must remain constant to support convolution operation and the training of stable models.

The second step is the normalisation of pixel values which converts the values in the image intensities to a common range. Normalisation enhances stability in training and hastens convergence in learning models. In order to improve the quality of the image, noise reduction methods are used to reduce the effects of camera noise, motion blur, and uneven lighting that is experienced in real-time shopping mall settings. In case of temporal modelling, frame sequencing is done by combining back-to-back facial frames in a fixed length sequence. These sequences are fed into the LSTM network and with this the system is able to learn emotional transitions as time goes on as opposed to using one frame predictions.

5. Deep Learning Architectures

5.1. CNN Architectures for Spatial Feature Extraction

The architecture of CNN consists of sequential convolutional layers (3 × 3 kernels) with ReLU activation and max-pooling. It has approximately 1.2M parameters. The Simple CNN was trained on FER-2013 for a maximum of 45 epochs with Adam optimizer and a batch size of 64. Mini-Xception model includes depth wise separable convolutions with residual connections, reducing parameter count to approximately 60K.

After pre-training, the final classification layer was removed from both architectures. Each CNN served as a fixed feature extractor, producing high-dimensional feature vectors for each input frame.

5.2. LSTM Architectures for Temporal Modelling

The pretrained CNN was then used to extract feature vectors from the consecutive facial frames to feed them as a 10-frame sequence into the LSTM part for temporal modelling.

In CNN- LSTM model, CNN-extracted spatial feature vectors from consecutive frames are passed to an LSTM layer for temporal modelling (~396K parameters for the temporal component). The LSTM evolves and learns temporal dependency among the sequence of frame-level facial features and the temporal representation is fed into a fully connected SoftMax layer for final emotion classification into seven classes.

The CNN–LSTM architecture consists of a fully connected layer with SoftMax activation for 7-class emotion classification. The CNN–LSTM model was trained with Adam optimizer with a learning rate of 0.0005 and batch size of 32. To prevent overfitting and the waste of training, early stopping was used with a patience of 15 epochs.

In this study, two different CNN architectures are employed for extracting spatial facial features, namely CNN and Mini-Xception, to analyse their effectiveness in facial emotion recognition. The CNN architecture serves as a baseline model and consists of a sequence of convolutional layers followed by ReLU activation and max-pooling operations. This architecture learns basic facial features such as edges, contours, and texture patterns from input images resized to 48 × 48 pixels. Despite its simplicity, the CNN provides valuable insights into how basic convolutional structures perform on facial emotion recognition tasks and establishes a reference for comparison with deeper architectures.

To improve feature representation and computational efficiency, the Mini-Xception architecture is also employed. Mini-Xception is a lightweight deep CNN model that uses depth-wise separable convolutions and residual connections to reduce the number of parameters while maintaining high recognition accuracy. This architecture enables the extraction of more discriminative and abstract facial features, such as subtle muscle movements around the eyes and mouth, which are critical for emotion recognition. Due to its efficiency and robustness, Mini-Xception is particularly suitable for real-time applications and resource-constrained environments, making it appropriate for deployment in smart shopping mall systems.

6. Classifiers

In the proposed framework, CNN and Mini-Xception are responsible for extracting frame-level spatial features, which are then organised into sequential feature vectors. These feature sequences serve as input to the CNN–LSTM architecture for temporal emotion modelling. The CNN–LSTM output is passed through a fully connected layer with SoftMax activation to classify emotions into seven categories: angry, disgust, fear, sad, happy, surprise, and neutral. This integration allows the framework to leverage both spatial and temporal information for facial emotion recognition and emotion-based shopping decision analysis.

Table 1. Consumer Behavioural Metrics

Emotion output	Psychological/consumer interpretation	Behavioural metric
Happy	Positive affect and satisfaction	High engagement; high purchase intention
Surprise	Elevated attention caused by an unexpected stimulus	High attention/engagement; potentially increased purchase intention
Neutral	Low or limited emotional response	Low–moderate engagement; uncertain purchase intention
Sad	Negative affect and reduced emotional satisfaction	Low engagement; low purchase intention
Angry	Strong negative affect, dissatisfaction or irritation	Low engagement; low purchase intention
Disgust	Strong aversion or rejection	Very low engagement; very low purchase intention
Fear	Perceived uncertainty, discomfort or threat	Low engagement; low purchase intention

Temporal emotion outputs were mapped to behavioural indicators based on environmental psychology literature. Sustained positive emotions (happiness, surprise) were associated with engagement and purchase intent; sustained negative emotions (sadness, anger, disgust) with disengagement.

7. Experimental Results and Discussions

7.1. CNN Performance on FER-2013

Table 1 presents Simple CNN and Mini-Xception classification performance on FER-2013.

Table 2. Performance of CNN models on FER-2013

Model	Accuracy	Precision	Recall	F1-Score	Parameters
Simple CNN	65.63%	65.78%	65.63%	65.47%	~1.2M
Mini-Xception	55.60%	56.12%	55.60%	54.89%	~60K

In the experimental setting of FER-2013, the Simple CNN obtained an accuracy of 65.63%, which is 10.03% better than Mini-Xception. Figure 3 and Figure 4 show Simple CNN training curves and confusion matrix. Figure 5 and Figure 6 show Mini-Xception and confusion matrix, respectively.

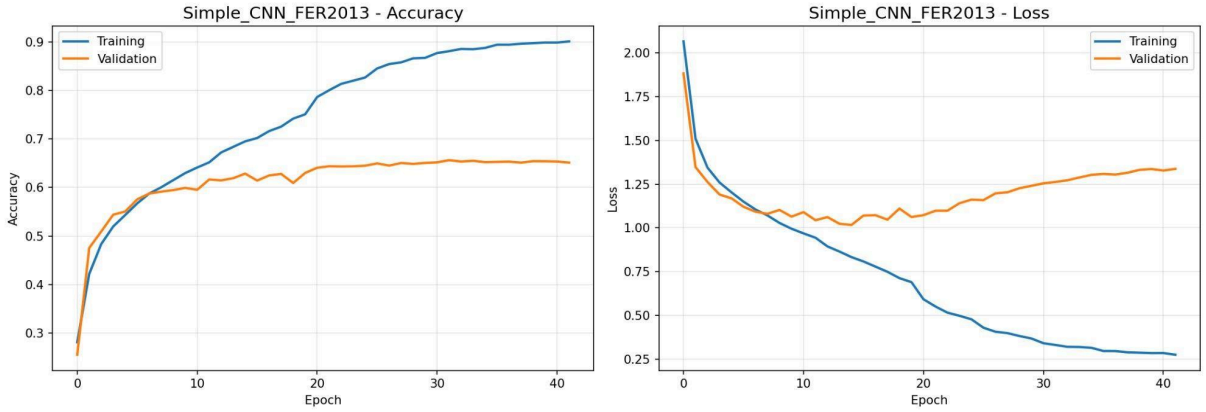

Figure 3. Simple CNN training curves on FER-2013. The accuracy of training was around 90% and the accuracy of validation was around 65%, indicating some overfitting of training. Convergence occurred by epoch 30.

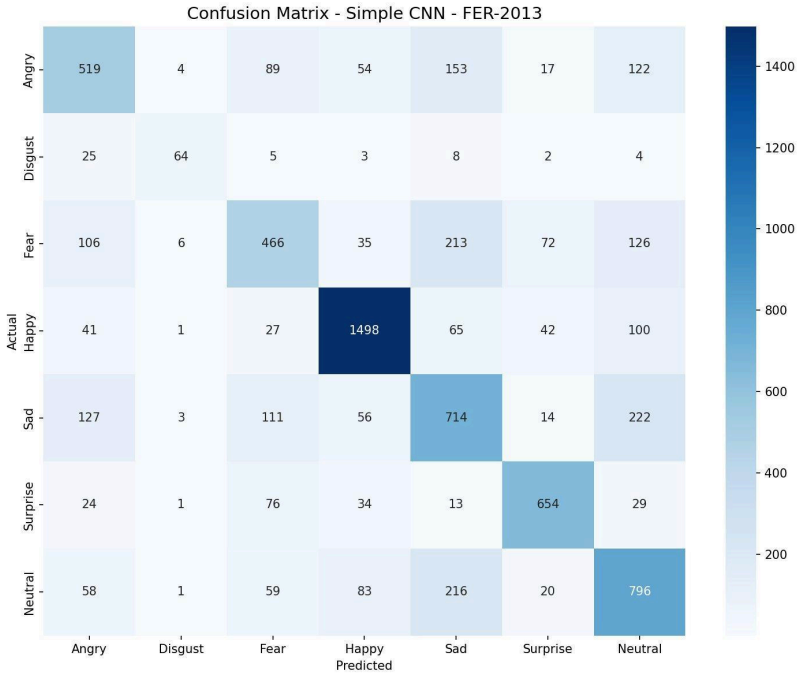

Figure 4. Simple CNN confusion matrix on FER-2013. Happy achieved highest recall (1,498/1,774 = 84.4%). The Fear class showed the lowest recall of the classes reported, with misclassification being more prevalent in the Angry, Sad and Neutral classes.

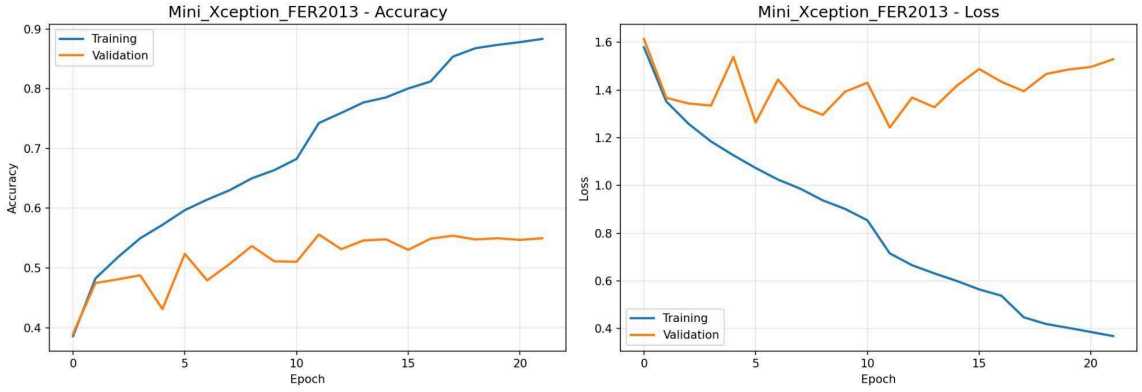

Figure 5. Mini-Xception training curves on FER-2013. Validation accuracy plateaued at ~55% while training accuracy continued rising, indicating both underfitting (insufficient model capacity) and overfitting simultaneously.

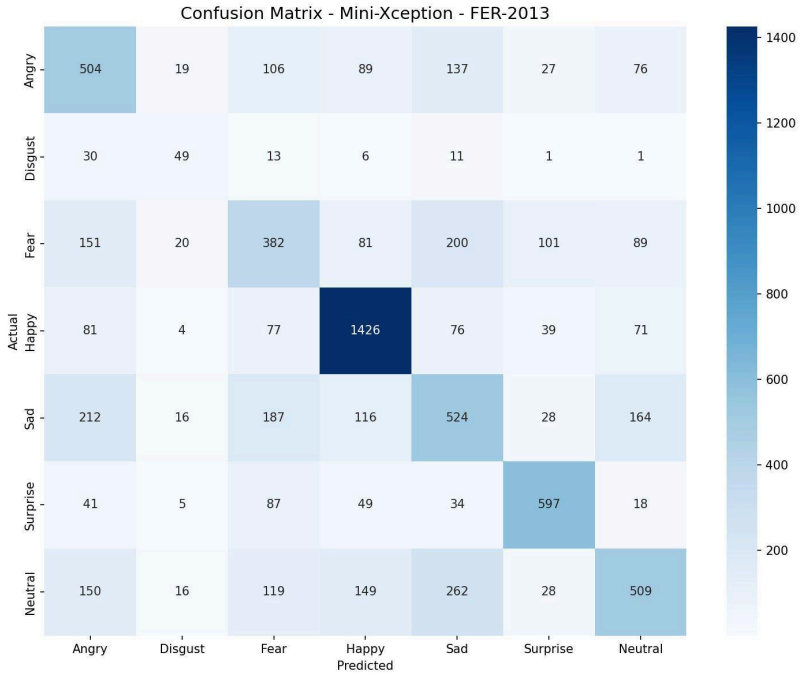

Figure 6. Mini-Xception confusion matrix on FER-2013. Performance degraded across all emotion classes relative to Simple CNN, with particularly poor discrimination between anger, fear, sadness, and neutral.

7.2. CNN-LSTM Performance on CK+

Table 2 presents LSTM performance on CK+ temporal sequences using pre-trained Simple CNN features.

Table 3: Performance of CNN–LSTM on CK+ temporal sequences.

Model	Accuracy	Precision	Recall	F1-Score	Parameters
CNN-LSTM	92.42%	93.94%	92.42%	91.79%	~2.05M

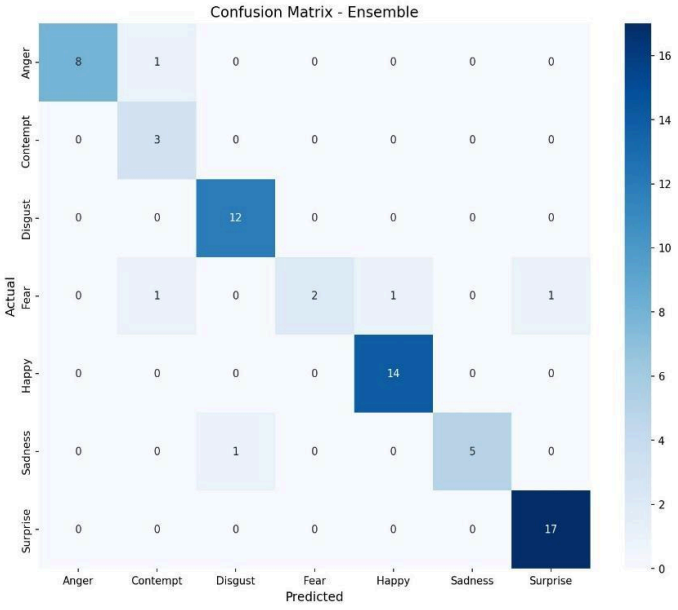

Figure 7. Ensemble confusion matrix combining all four LSTM models. Accuracy (92.42%) did not exceed the best individual model, indicating that error patterns across architectures were subsets rather than complements.

The accuracy of CNN–LSTM reported on CK+ is not to be interpreted as a statement of the superiority of temporal modelling over spatial modelling, as the CNN and CNN–LSTM experiments were tested on different datasets and data nature.

7.3. Temporal vs. Static Comparison

Table 4. Impression of the experimental results for all modelling settings.

Model	Dataset	Type	Accuracy	Finding
Simple CNN	FER-2013	Spatial	65.63%	Competitive baseline
Mini-Xception	FER-2013	Spatial	55.60%	Underfitting (high bias)
CNN-LSTM	CK+	Temporal	92.42%	No gain over best single

The reported accuracy values are not directly related to the progress of the spatial to temporal modelling but were obtained under various settings of the dataset and the task.

The results, on the other hand, show the performance of each modelling strategy on the respective experimental dataset: the Simple CNN and Mini-Xception models were tested on FER-2013 to classify the spatial emotions, while CNN–LSTM was tested on CK+ to classify the temporal sequence emotions.

The gap of 10.03% points between Simple CNN and Mini-Xception on FER-2013 could be related to the model capacity and representation efficiency. Mini-Xception's depthwise separable convolutions reduce parameter count by approximately 20× (~60K vs. ~1.2M), introducing higher bias. With 35,887 training images, FER-2013 provides sufficient data to support Simple CNN's larger capacity without severe overfitting. Mini-Xception's efficiency-oriented design, appropriate for deployment, is suboptimal for pre-training on this dataset size.

While the FER-2013 archive mainly includes images of individual faces, the CK+ archive consists of temporally ordered expression sequences. So, it is not only temporal modelling that affects accuracy, since the difference was observed. CK+ contains only 261 static training samples, insufficient for CNN spatial learning, while LSTM temporal augmentation (overlapping sub-sequences) effectively triples available training data. Second, information content: CK+ expression sequences progress from neutral to peak intensity; the temporal trajectory provides discriminative information absent from any single frame. A single frame showing partial mouth opening is ambiguous; the trajectory from closed to open disambiguates surprise from happiness.

The two-stage approach takes advantage of the complementary qualities of FER-2013 and CK+. FER-2013 is a relatively large database of facial images suitable for spatial representation learning, while CK+ is temporally ordered and annotated expression sequences suitable for temporal modelling. These complementary properties also help to motivate the two-stage experimental design, but also render direct comparison of the accuracy values obtained impossible. FER-2013 provides spatial diversity (35,887 images, uncontrolled conditions) but no temporal structure. CK+ provides temporal sequences (981 annotated progressions) but limited spatial diversity (123 subjects, controlled conditions). Partitioning spatial learning to stage one and temporal learning to stage two prevents either dataset from compensating for the other's limitations.

Table 5 characterises the domain gap across many dimensions.

There is a key difference between the two datasets with regard to expression intensity. CK+ contains posed expressions progressing to full intensity; retail customers exhibit subdued spontaneous expressions. The mutual information between subdued expressions and emotion categories is lower than between peak expressions and categories. This difference points to the difficulty in transferring performance from the standardised benchmark expressions to spontaneous expressions in real retail contexts.

Table 5. Distribution Characteristics of FER-2013 and CK+

Dimension	FER-2013	CK+
Resolution	48×48 grey	640×490 colour
Compression	JPEG	Uncompressed
Expression	Spontaneous+posed	Posed only
Intensity	Variable	Neutral→peak
Illumination	Uncontrolled	Studio controlled
Viewing angle	Mostly frontal	Frontal only
Demographics	Diverse	Mostly Caucasian
Occlusion	Minimal	None
Background	Cropped	Uniform

8. Limitations

This work has been tested on an online database. However the challenges may increase when implemented on a real time shopping mall dataset.

9. Conclusion

In this paper, a CNN–LSTM model was proposed for the recognition of facial emotions for future emotion-aware retail analytics. In the two-stage approach, the spatial representation learning and the temporal sequence modelling are carried out with the complementary properties of FER-2013 and CK+. The following experiments should be regarded within the context of the limitations of the benchmark datasets. The study in particular does not validate purchases made by the customer, sales results or actual retail behaviour directly. Hence, the facial-emotion recognition performance observed in this study alone cannot draw a cause and effect or predictive relationship between facial emotions and purchasing decision. Future work will address the identified limitations through: (1) Collection of consented retail expression datasets capturing spontaneous emotions; (2) Development of retail-specific emotion taxonomies aligned with consumer behaviour theory; (3) Validation of pattern of facial emotions with the actual shopping outcome with controlled experimental designs to test whether the facial-emotion pattern correlates with the actual outcome of the shopping; (4) Domain adaptation techniques to bridge the laboratory–retail distribution gap; (5) Pilot deployment with partner retailers to evaluate real-world performance and privacy architecture effectiveness.

References

Boz, H., & Kose, U. (2018). Emotion extraction from facial expressions by using artificial intelligence techniques. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 9*(1), 5–16. https://doi.org/10.70594/brain/v9.i1/1

Deshmukh, A., & Patil, H. (2026). Emotion recognition method based on convolutional neural network and Black Widow optimization algorithm. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 17*(1), 26–39. https://doi.org/10.70594/brain/17.½

Fu, Y., Wu, X., Li, X., Pan, Z., & Luo, D. (2020). Semantic neighborhood-aware deep facial expression recognition. *IEEE Transactions on Image Processing, 29*, 6535–6548. https://doi.org/10.1109/TIP.2020.2991510

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735Sariyanidi, E., Gunes, H., & Cavallaro, A. (2015). Automatic analysis of facial affect: A survey of registration, representation, and

recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(6), 1113–1133. https://doi.org/10.1109/TPAMI.2014.2366127

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition and Emotion, 24*(8), 1377–1388. https://doi.org/10.1080/02699930903485076

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2584–2593). IEEE. https://doi.org/10.1109/CVPR.2017.277

Li, S., & Deng, W. (2019). Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing, 28*(1), 356–370. https://doi.org/10.1109/TIP.2018.2868382

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Lopes, A. T., de Aguiar, E., De Souza, A. F., & Oliveira-Santos, T. (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition, 61*, 610–628. https://doi.org/10.1016/j.patcog.2016.07.026

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 94–101). IEEE. https://doi.org/10.1109/CVPRW.2010.5543262

Mellouk, W., & Handouzi, W. (2020). Facial emotion recognition using deep learning: Review and insights. *Procedia Computer Science, 175*, 689–694. https://doi.org/10.1016/j.procs.2020.07.101

Mohammadpour, M., Khaliliardali, H., Hashemi, S. M. R., & AlyanNezhadi, M. M. (2017). Facial emotion recognition using deep convolutional networks. In *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)* (pp. 17–21). IEEE. https://doi.org/10.1109/KBEI.2017.8324974

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10). IEEE. https://doi.org/10.1109/WACV.2016.7477450

Rouast, P. V., Adam, M. T. P., & Chiong, R. (2021). Deep learning for human affect recognition: Insights and new developments. *IEEE Transactions on Affective Computing, 12*(2), 524–543. https://doi.org/10.1109/TAFFC.2018.2890471

Velammal, S., Leo, A., Kuncha, R. K., & Rani, M. (2025). Analysis of consumer decision on e-commerce platforms with video advertising and brain data. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 16*(4), 225–241. https://doi.org/10.70594/brain/16.4/13

Yu, Z., Liu, G., Liu, Q., & Deng, J. (2018). Spatio-temporal convolutional features with nested LSTM for facial expression recognition. *Neurocomputing, 317*, 50–57. https://doi.org/10.1016/j.neucom.2018.07.028

Zarezadeh, E., & Rezaeian, M. (2016). Micro expression recognition using the Eulerian video magnification method. *BRAIN: Broad Research in Artificial Intelligence and Neuroscience, 7*(3), 43–54. https://doi.org/10.5281/zenodo.1044975