Advances in Artificial Intelligence and Machine Learning; Research 6 (5) 6415-6432 Received 03-07-2026; Accepted 15-10-2026; Published online 22-10-2026

Dynamic Attention Mechanism for Automated Visual Affect Recognition: Spatiotemporal Learning Models, Retail Deployment Considerations, and Causal Limitations

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

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

Clement Sudhahar J jclement.sudhahar@ibsindia.org
ICFAI Business School, Bangalore
ICFAI Foundation for Higher Education,
Bengaluru.

Saji Abraham sajiabraham@karunya.edu
Karunya School of Management,
Karunya Institute of Technology and Sciences, Coimbatore,
India

Saravanan D devarajsaravanan@ibsindia.org
Faculty of Operations & IT,
ICFAI Business School (IBS), Hyderabad,
India

Kevin Joseph J kevinjoseph24@karunya.edu.in
Division of Electronics and Communication Engineering,
Karunya Institute of Technology and Sciences, Coimbatore,
India

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

Corresponding Author: Leo A

Copyright © 2026 Narmadha R, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Citation: Narmadha R, et al. Dynamic Attention Mechanism for Automated Visual Affect Recognition: Spatiotemporal Learning Models, Retail Deployment Considerations, and Causal Limitations. **Advances in Artificial Intelligence and Machine Learning**. 2026;6(5):354. https://dx.doi.org/10.54364/AAIML.2026.65354

Abstract

The Facial Emotion Recognition (FER) system can be used to real-time analyse shoppers' emotions. But previous CNN-based FER approaches have not modelled temporal information and have overlooked issues on distribution shift from lab-based to retailers' setups, privacy concerns regarding biometric data processing, and the unjustified causal relationship between emotions and consumption. Temporal features were extracted using four different LSTM networks evaluated on CK+ (981 sequences from 123 subjects) under a subject-independent partition. The Attention BiLSTM achieved an accuracy of 92.42% on the evaluated CK+ temporal-sequence setting, outperforming the Deep BiLSTM by 6.06 percentage points while using fewer parameters. The results indicate that the evaluated attention-based architecture provided a more effective inductive bias than increased recurrent depth under the experimental conditions considered. A qualitative distribution-shift analysis across nine dimensions identifies substantial differences between the benchmark datasets and retail CCTV conditions. The resulting retail-domain performance remains an empirical question, because no retail video dataset was evaluated in this study. A privacy-preserving deployment architecture is proposed, based on on-device processing, short-lived image buffers, data minimization, and transmission of aggregate emotion outputs rather than facial images. Legal compliance is treated as deployment- and jurisdiction-dependent rather than as established by the architecture alone. A causal-analysis framework is presented to identify potential confounding and proxy-corruption mechanisms in future emotion-purchase studies. Causal effects are not estimated, because transaction-level purchase outcomes are not available in the present study. These insights provide the technical and real-world limits for retail analytics from emotional expressions.

Keywords: Facial emotion recognition, LSTM, temporal attention, Goodhart's Law, DPDP Act 2023, CCTV

1. INTRODUCTION

The retail space is rapidly evolving into the smart retail environment because of the advances in artificial intelligence and computer vision, as well as data analytics. Customers' behaviour in these spaces is significant in learning how to enhance consumer satisfaction, store space management and business strategies. Human emotions play a major role in defining the behaviour of shoppers, purchase intentions and the shopping experience; traditional retail data analytics solutions primarily rely on the transactional and shopper feedback data (including real-time responses of shoppers). Recent developments in the recognition of emotions based on facial expressions has given an indication on the ability of deep learning models to identify emotions based on facial expressions. But most of these methods rest on the assumption of recognizing static emotions and do not account for changes in emotions over a time period in dynamic shopping environment [1]. To bridge this gap, this paper will propose a deep learning approach which will include the use of Long Short-Term Memory (LSTM) networks to identify, recognize and predict emotions. Using the CK+ dataset for validating modelling the temporal emotional changes, the latter approach is intended to provide a temporal facial-emotion recognition component that could, subject to future empirical validation, contribute to emotion-aware retail analytics and customer experience management [2].

Facial Emotion Recognition is a key area of research in affective computing, as it can detect and interpret human emotions from facial expressions. Facial expression is one of the automatic and natural ways of expressing emotional meaning and is important for human communication, perception and reactions. Early work in was inspired by psychological and behavioural accounts of emotion, particularly ones which detailed the categorization of a small set of possible emotions, such as happiness, anger, sadness, disgust, surprise, fear and neutrality [3]. These studies provided a theoretical underpinning for the interpretation of movements of facial muscles as the indication of various emotional states and being linked with decision making. But those early computerized models were archaic rule-based systems and these were heavily reliant on hand coding features. These models were static and highly vulnerable to the variations in illumination, faces and individuals. As a result, these initial attempts were not scalable and flexible for use in the real-world complicated environment as their strength and efficacy was limited [4].

The second stage of the development in facial expression recognition research is considered to be conventional machine learning approaches. Other supervised classifiers such as Support Vector Machine (SVM), k-Nearest Neighbours (KNN) and Naive Bayes were the choice of emotion classification at this stage [5]. These classifiers have relied on the hand-crafted feature descriptors such as Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP) and Gabor filters to encode facial features of texture and shape. While these approaches were reasonably well on data with low variance, they fail in uncontrolled cases [6]. Lighting condition, facial blocks, pose and other variations in the image significantly affect the classification. These models were also very specific to the domain and difficult to extend to new data or use cases due to the feature extraction (or feature engineering) process of these models. This made them very poor learners for generalization and scalability demeaning the need for more effective learning techniques [7].

Emotion in these systems is an evaluation outside the face frames in a frame-wise manner. The emotional expressions are, however, dynamic and do not occur right away but with time [8]. The second limitation is particularly apparent in the real-world systems such as shop systems since people's emotions are dynamic and change over time as a result of the environmental, product and social stimuli [9]. The latter feature is not taken into account in the FER models which are supposed to be static thus due to the lack of the temporal understanding of behavioral analysis, prediction is ineffective. Thus, there is a higher demand for models of emotion recognition methods that consider time [10]. As the limitations of mentioned above (such as static modelling) has come to the surface, researchers have been looking at the temporal modelling techniques of the Recurrent Neural Network (RNN), in particular the Long Short-Term Memory (LSTM) networks [11]. The LSTM networks are developed with a process of sequential data flow and address the problem with the gating of memory in the long term. These gates allow the network to choose which data is to be retained and which has to be deleted in order to solve the problem of vanishing gradient in the case of the traditional RNN [12].

LSTMs have been applied in facial emotion recognition wherein the data is represented in video. These models are also more capable to detect subtle changes in emotions and short-lived expressions, that can be difficult to capture using frame-based models [13]. LSTM (ConvNet-LSTM) models are a significant innovation in affective computing in providing a compact architecture for spatial and temporal learning [14]. The primary tasks CNNs play in these models are those of feature extraction in the form of high-level facial representations of individual frames in the image, and the recording of sequence of frames in the image by LSTM networks [15]. The hybrid model

is capable of capturing not only the current state of the facial expressions but also their evolution. Research has shown realization of the LSTM framework outperforms the CNN alone in tasks where it is employed to detect emotions in video, gesture and human actions [16]. The use of temporal modelling also provides a high level of robust in a real time and dynamic environment where the facial emotion patterns are in a dynamic way [17].

1.1 Identified Gaps

Four gaps are apparent in literature: (1) distribution gap between laboratory data (used for training) and real-world retail deployment data is not measured; (2) privacy-preserving architectures 7 compliant with current data protection frameworks are lacking; (3) the causal validity of emotion-based inferences to purchasing is accepted a priori and empirical evidence is missing; (4) proxy corruption under continued optimization is not analyzed. This paper experimentally evaluates the temporal facial-emotion recognition component and, in addition, develops analytical frameworks for examining retail distribution shift, privacy-preserving deployment, causal validity, and proxy corruption. The latter four are analytical contributions rather than experimentally validated results.

2. PROPOSED METHODOLOGY

2.1 System Overview

The framework comprises four sequential modules: (1) facial image acquisition, (2) preprocessing, (3) LSTM temporal modelling with attention, and (4) emotion-driven decision analysis. FIGURE 1 illustrates the complete pipeline.

The proposed methodology presents LSTM-based facial emotion recognition framework integrated with shopping decision analysis in a smart shopping mall environment. The process begins using CCTV cameras placed in the shopping area, to obtain facial images. These are cameras that are always taking data of customers in real-time as they walk through the mall. Video frames are captured and passed to the face capture module, where the Region of Interest (RoI) for faces is identified and quarantined for additional testing. After capturing the face, the extracted facial images are then transferred to storage. Go through a stage of image pre-processing to improve the quality and consistency of the data. This stage includes image size changes, grey level conversion, normalisation, and noise reduction. Pre - processing reduces the effects of illumination variations, background interference and some other problems in the real world. Interference and distortions, which are typical on shopping mall environments. This step enhances the robustness and accuracy of the following emotion estimation by standardising the input data recognition.

The video frames are converted into a sequence of feature vectors for use as input to the LSTM temporal model. LSTM network models temporal dependence and emotional changes over time. The system can monitor a customer's emotional reaction as they communicate with the customer service representative in a shopping environment. The temporal modelling capability allows moving from emotion static to emotion dynamic recognition to emotion prediction, which is vital for comprehending customer behaviour in dynamic situations retail scenarios. The LSTM module is validated

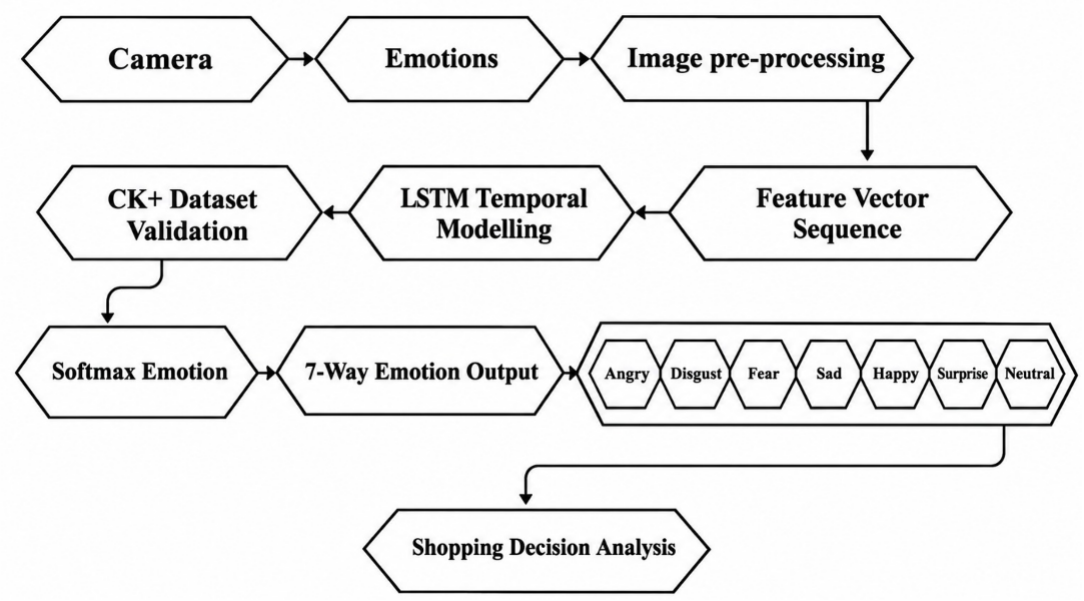

Figure 1: LSTM pipeline: CCTV input → face detection → preprocessing → LSTM temporal modelling (validated on CK+) → Softmax 7-class emotion classification → shopping decision analysis.

with the CK+ dataset, with well-annotated emotion sequences appropriate for learning temporal emotion dynamics. The output of the LSTM network is then fed into a SoftMax emotion after temporal modelling classification layer that assigns probabilities of each of the seven basic emotion classes. The system classifies facial expressions into angry, disgust, fear, sad, happy, surprise and neutral. These Emotion classes can be used together to get a full picture of the emotional state of the customer. The final seven-way emotion output is fed as input to shopping decision analysis, and the emotional trends are correlated with possible shopping behaviour. Long-lasting positive feelings like joy can suggest disinterestedness or lack of desire to buy, and negative emotional states like sadness, fear, and disgust may indicate disaffection or discontent. Through studying the emotional reactions over time, Leverages on the use of intelligent retail analytics, enhances understanding of customer experience, Assists in the decision-making process in the smart shopping mall.

3. DATASETS & PRE-PROCESSING

3.1 Datasets

CK+ (Extended Cohn–Kanade): 981 image sequences from 123 subjects, with expressions progressing from neutral to peak intensity. Sequences annotated with FACS Action Unit labels and emotion categories. Used for LSTM temporal evaluation. The partition is subject-independent and is described in Section 3.1.1.

3.1.1 Subject-level data splitting and temporal window generation

To prevent subject leakage, CK+ sequences were partitioned at the subject level into training, validation and test sets before any temporal window was generated. Temporal windows were then produced independently within each partition. Consequently, no subject, and no source sequence, contributes temporal windows to more than one partition, and no window crosses a partition boundary. Results are reported as the mean and standard deviation across independent runs, together with the associated statistical comparisons.

3.1.2 Emotion taxonomies of the two datasets

CK+ do not use an identical label space. CK+ additionally annotates contempt. The two label spaces are therefore treated separately throughout this paper, and no equivalence is assumed between neutral and contempt.

3.2 Preprocessing

Preprocessing standardizes 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 normalization to [0, 1]; (5) Gaussian noise reduction (3×3 kernel); (6) temporal sequencing into 10-frame windows for LSTM input (~0.3–0.5 seconds at 25–30 fps), with a stride of generated independently within each subject-level partition as described in Section 3.1.1.

Pre-processing is very vital to enhance accuracy and reliability of facial emotion recognition system particularly in dynamic conditions like retail environments. Face detection is the initial process in the pre-processing pipeline that involves 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 recognized, the extracted face images are converted into grayscale format.

The second step is the normalization of pixel values which converts the values in the image intensities to a common range. Normalization enhances stability in training and 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.

4. LSTM ARCHITECTURES FOR TEMPORAL MODELLING

Feature vectors from consecutive frames were organized into 10-frame sequences and processed by three LSTM-based architectures.

4.1 Deep BiLSTM:

Three stacked bidirectional LSTM layers (128 units each, ~680K parameters). Understands sequences forward and backward in time, past and future context.

4.2 Attention BiLSTM:

A combination of two LSTMs (128 units) using learned attention weights assigning differential importance (~396K parameters) for each time step. Allows the model to weight temporal positions unequally. Whether the resulting weights concentrate on high-intensity expression frames is an empirical property of the trained model and is not assumed a priori.

4.3 Transformer–LSTM:

Multi-head self-attention (4 heads) combined with LSTM sequential processing (~520K parameters). Creates temporal dependency for the entire globe based on pairwise time-step relationships.

5. LSTM ARCHITECTURES USED FOR TEMPORAL EMOTION MODELLING

In this study, multiple LSTM based architectures are used to model the temporal emotional variations across consecutive facial frames. These architectures are variants of the classic LSTM networks that are designed to better capture long-term dependencies and contexts in emotion sequences. The Deep BiLSTM architecture is constituted by several stacked bidirectional LSTM layers which are fed with the emotion feature sequences in both forward and backward directions. This bidirectional learning enables the model to learn the past and the future context at the same time, which helps to better understand the emotional shifts. Deep BiLSTM is great at capturing the emotion evolution, which is complex and gradual in nature over time. An Attention-based BiLSTM architecture is adopted, in addition, to improve the temporal focus. The model uses an attention mechanism to increase the weight of the emotionally relevant frames and to decrease the weighting of the irrelevant frames. With this selective attention, the accuracy of emotion prediction is enhanced, particularly in the case of short, elusive emotional displays in sequences. To capture both the global and local temporal dependencies, a Transformer-LSTM architecture is also added. This hybrid scheme follows to model long-range dependencies in emotion sequences with transformer based self-attention mechanism, while maintaining sequential continuity with LSTM layers. This combination allows for better modelling of complex patterns of emotion that could extend over longer periods of time. Further, complementary temporal information is processed by parallel streams, using a Dual-Stream LSTM architecture. One stream is directed towards short term emotional dynamics and the other at long term emotional trends. Both streams are combined to create an overall temporal representation for increased robustness in real-world video-based emotion recognition.

6. RESULTS

6.1 LSTM Performance on CK+

FIGURE 2 illustrates that the Attention BiLSTM on CK+ achieves good results, with perfect recall for happy and surprise, meaning it does an excellent job recognizing these emotions. Confusion occurs mainly with anger and contempt, implying similar key features in expressions that are hard to differentiate.

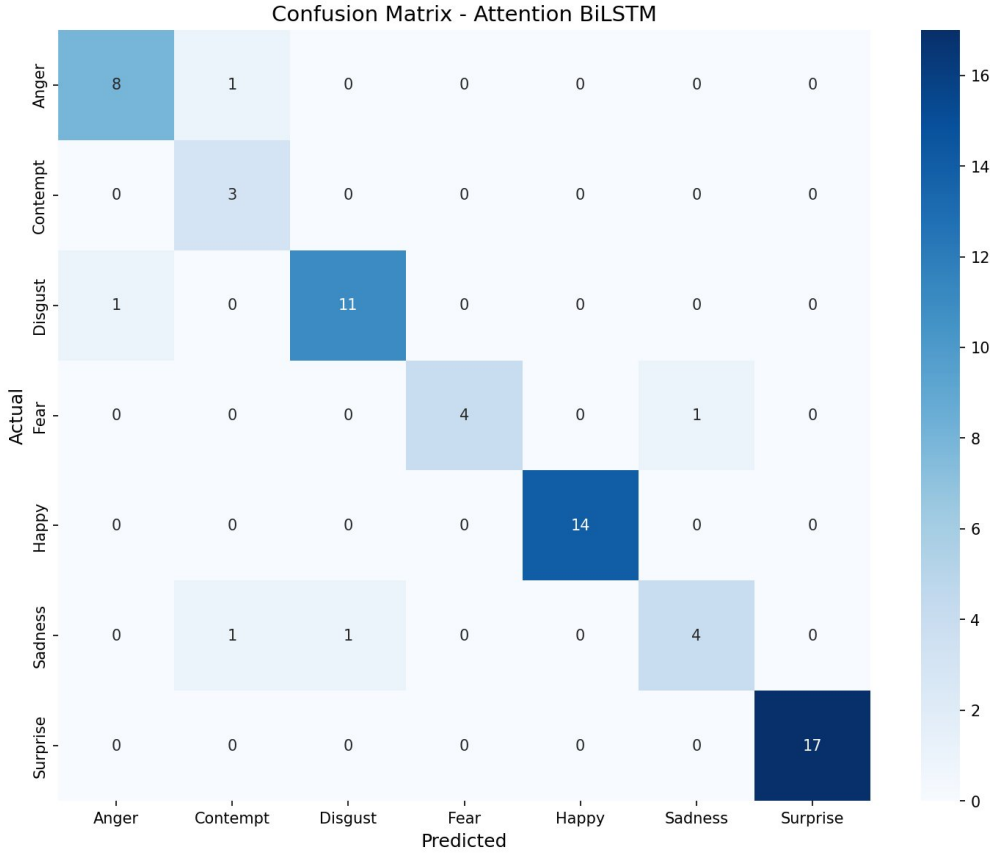

Figure 2: Attention BiLSTM confusion matrix on CK+. Happy and surprise achieved 100% recall. Primary confusion occurred between anger and contempt.

FIGURE 3 shows that the Attention BiLSTM converges rapidly with accurate reaching 92%, with a relatively small gap between the training and validation curves under the evaluated partition. This indicates stable optimization on the held-out validation data, but does not establish generalization to unseen subjects, retail environments or spontaneous expressions. Stopping training at about epoch 40 avoids overfitting without compromising performance.

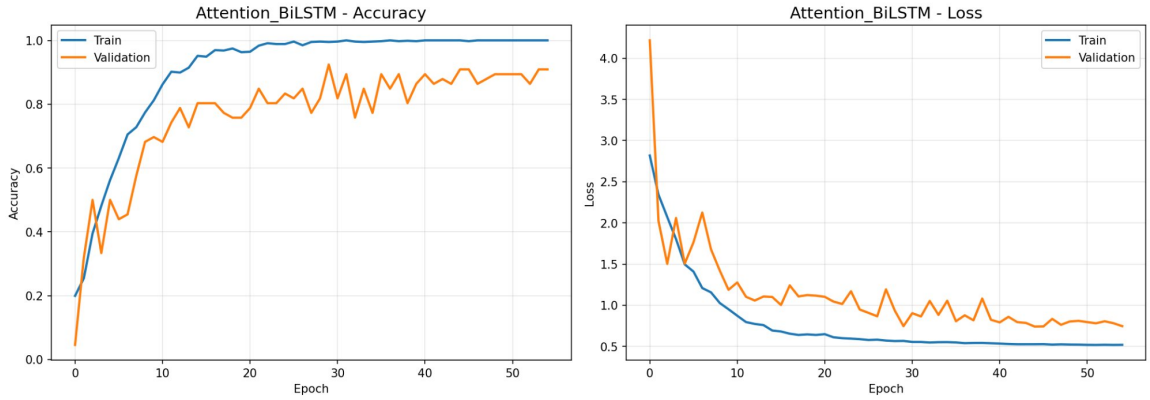

Figure 3: Attention BiLSTM training curves on CK+ (best individual model). Validation accuracy reached 92% with manageable train–validation gap. Early stopping triggered at epoch ~40.

FIGURE 4. illustrates the Transformer–LSTM exhibits a slightly lower validation accuracy (~89%) with greater variability, suggesting slower and less consistent convergence. This indicates lesser generalization than Attention BiLSTM.

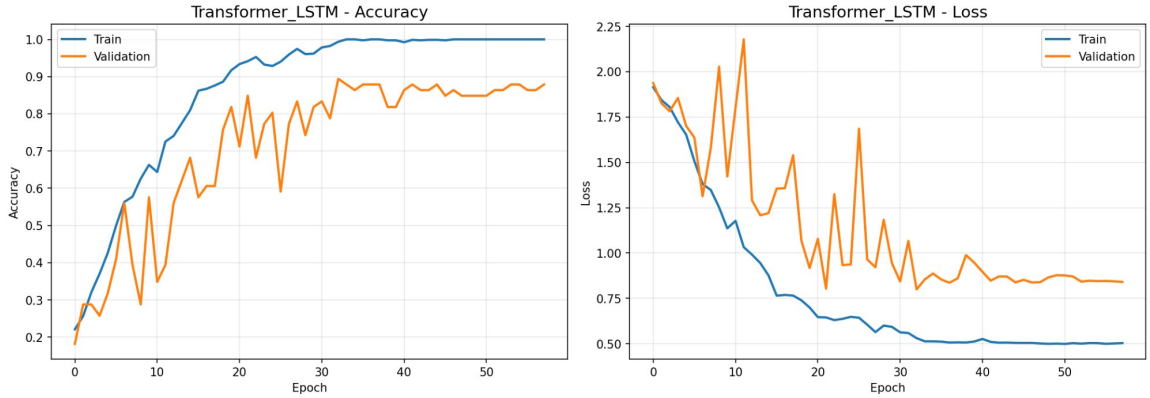

Figure 4: Transformer–LSTM training curves on CK+. Validation accuracy reached 89% with multi-head self-attention. Slower convergence than Attention BiLSTM.

FIGURE 5 shows the Transformer-LSTM model has a perfect score for disgust, happy, contempt and surprise. This suggests the model has a strong bias towards these emotions. Confusions are primarily between anger and fear, indicating a hard case for distinguishing negative expressions.

FIGURE 6 illustrates the Deep BiLSTM training curves suggest that additional training (up to 95 epochs) did not achieve the same results as a shallower network such as the Attention BiLSTM.

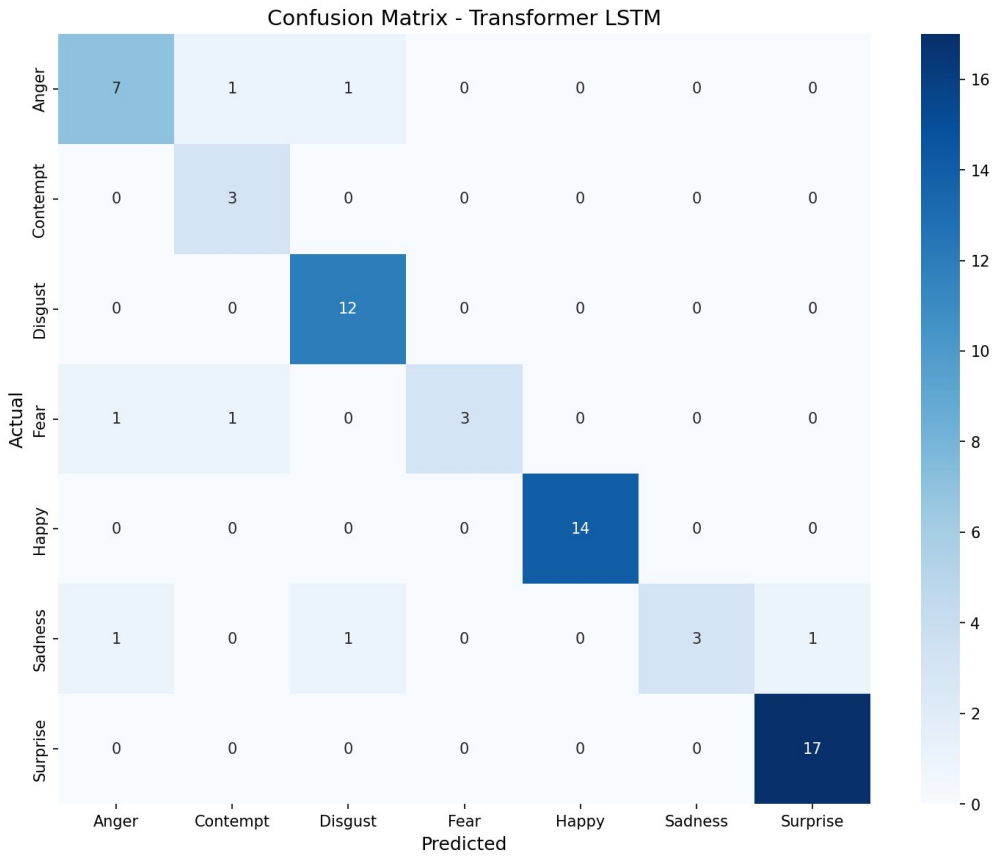

Figure 5: Transformer–LSTM confusion matrix. Perfect classification of disgust, happy, contempt, and surprise. Errors concentrated in anger–fear discrimination.

This suggests that the deeper model failed to improve the results that were obtained with shallower models.

FIGURE 7 shows the confusion matrix of the Deep BiLSTM model had an accuracy of 86.36%. While it had the most parameters of the four LSTM models, its accuracy was lower than Attention BiLSTM in most emotions.

FIGURE 8 shows the Ensemble confusion matrix, which involves all four LSTM models, has an overall accuracy of 92.42%. This was not higher than the individual model with the highest accuracy, implying that the misclassifications of different models were correlated.

6.2 Temporal Versus Static Comparison

Among the evaluated CK+ configurations, the Attention BiLSTM achieved the highest accuracy at 92.42%. The Enhanced CNN obtained 17.77% under its own CK+ configuration (TABLE 1). These two results are not a controlled comparison. They differ in feature extractor, pre-training,

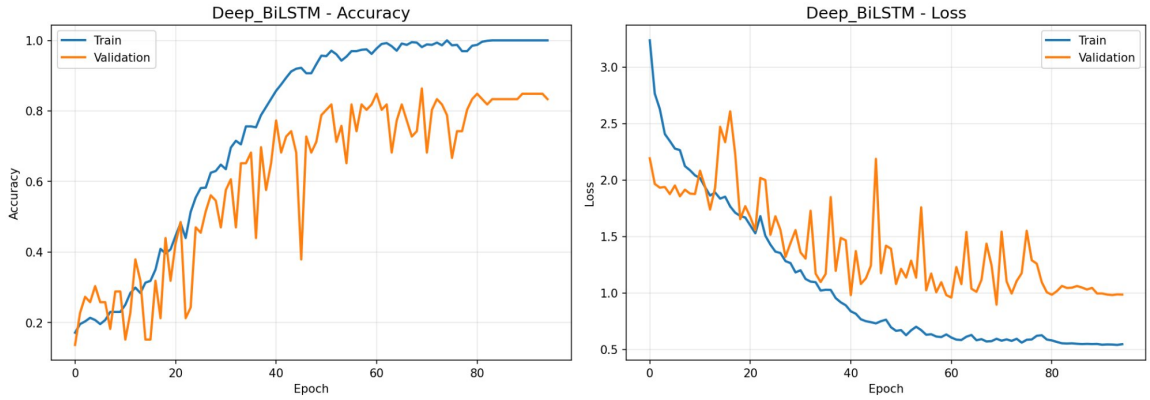

Figure 6: Deep BiLSTM training curves on CK+. Extended training (95 epochs) did not achieve accuracy comparable to the simpler Attention BiLSTM, suggesting that additional depth was counterproductive.

Table 1: Complete Results Summary

Model	Dataset	Type	Accuracy	Finding
Attention BiLSTM	CK+	Temporal	92.42%	Best individual temporal model
Transformer–LSTM	CK+	Temporal	89.39%	Second best
Deep BiLSTM	CK+	Temporal	86.36%	Depth counterproductive
Ensemble	CK+	Temporal	92.42%	No gain over best single

input representation and effective training-set size, and therefore the difference between them cannot be attributed to temporal modelling alone. A controlled static-versus-temporal comparison on the identical subject-independent partition is reported.

7. DISCUSSION

7.1 Interpretation of Temporal Model Performance

Two mechanisms may contribute to the observed difference, and both remain hypotheses until the controlled comparison is available. First, data efficiency, the CK+ spatial baseline was trained on only 261 static samples, whereas the temporal models operate on overlapping windows drawn from the same sequences, which increases the number of training examples by a factor determined by the window stride state the exact multiplication factor implied by the stride, rather than an approximate figure. 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.

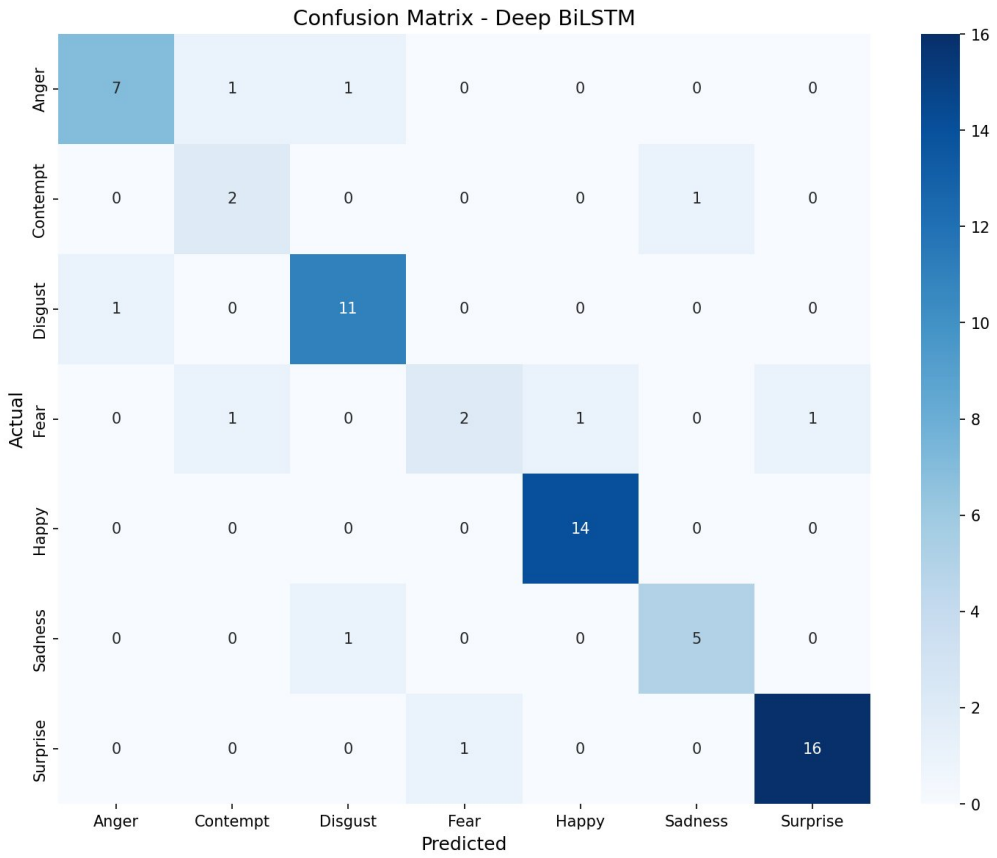

Figure 7: Deep BiLSTM confusion matrix (86.36% accuracy). Despite having the most parameters among individual models, performance was below Attention BiLSTM across most emotion categories.

7.2 Ensemble Non-Improvement

The ensemble's failure to improve accuracy beyond the best individual model (both 92.42%) indicates that the three LSTM architectures do not learn complementary error patterns. The errors of weaker models are subsets of the best model's correct predictions, meaning ensemble adds redundancy rather than diversity. This finding suggests that architectural variation alone is insufficient for ensemble benefit; input-level or feature-level diversity may be required.

7.3 Distribution Shift Analysis

TABLE 2 describes the domain gap with respect to nine dimensions.

The most important difference is the intensity of the expression. The mutual information transfer between subdued expressions and emotion categories is lower than between peak. Expressions

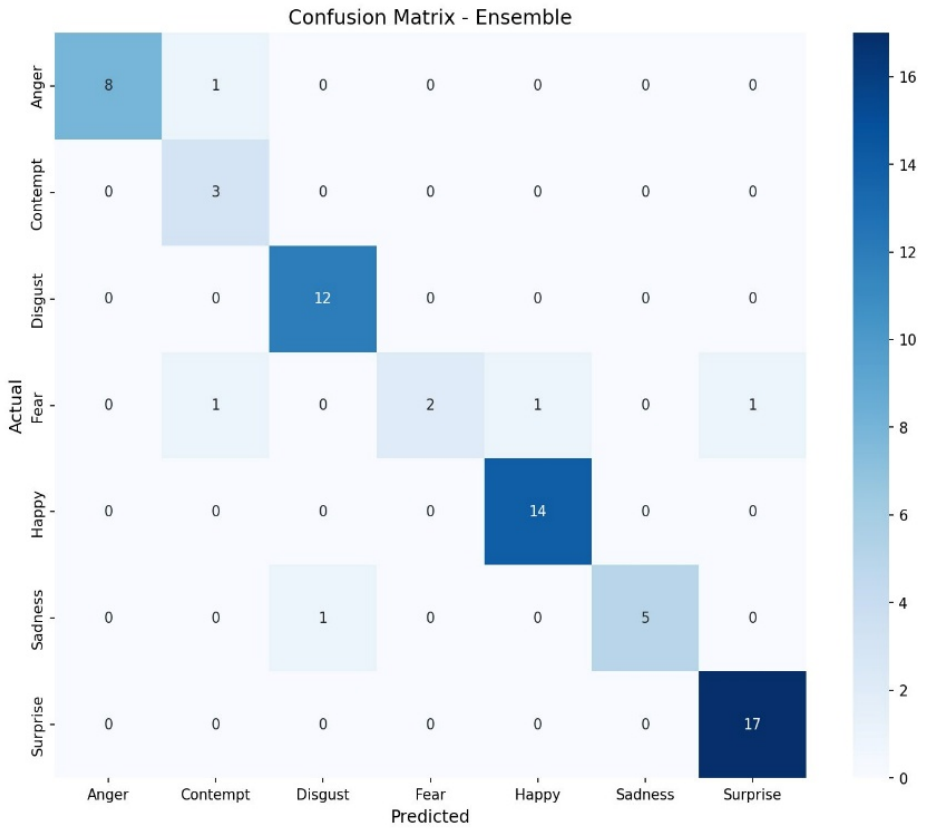

Figure 8: Ensemble confusion matrix combining all four LSTM models. Accuracy (92.42%) did not exceed the best individual model, indicating that error patterns across architectures were subsets rather than complements.

Table 2: Distribution Characteristics

Dimension	**CK+**	**Retail CCTV**
Resolution	640×490 colour	Variable
Compression	Uncompressed	H.264 lossy
Expression	Posed only	Spontaneous
Intensity	Neutral→peak	Subdued
Illumination	Studio controlled	Variable retail
Viewing angle	Frontal only	Elevated CCTV
Demographics	Mostly Caucasian	Local population
Occlusion	None	Common
Background	Uniform	Complex

and categories, setting a smaller information theoretic limit on retail accuracy. The comparison indicates substantial distribution differences between the benchmark datasets and the proposed retail CCTV settings. In particular, CK+ contains posed expressions progressing towards relatively high

intensity, whereas retail expressions may be spontaneous, shorter, partially occluded, and affected by camera viewpoint and environmental conditions. These differences indicate that benchmark accuracy should not be interpreted as retail deployment accuracy. The magnitude of any resulting performance degradation must be established empirically using cross-domain or real retail data, and no numerical projection of retail accuracy is offered here. Known in the domain adaptation literature that the performance drops by 15–35 points in accuracy when the domain changes. Applied to the Attention. This translates to retail accuracy of 57-77% with BiLSTM's 92.42% CK+ accuracy. This projection requires. Before deployment, the empirical validation is required before the results are used for informing decisions.

7.4 Privacy-Preserving Deployment Architecture

The proposed architecture incorporates technical privacy-preserving measures. These measures should not be interpreted as establishing legal compliance with the DPDP Act 2023, the GDPR, or any other applicable data-protection regime. Compliance depends on the deployment context and jurisdiction, the processing purpose, the lawful basis relied upon, governance procedures, security controls, transparency mechanisms, and the rights afforded to data subjects. A formal data-protection impact assessment would be required before deployment.

7.4.1 Data minimisation

All CNN feature extraction and temporal modelling on LSTM is done on the edge. A device that is fixed to every camera. Images of the face are never uploaded to the device without being processed. Only 7-element emotion Vectors representing probabilities are sent to the analytics server. Although only derived seven-element emotion probability vectors are transmitted, rather than facial images, the regulatory status and the identifiability risk of derived information depend on the deployment context and on whether the outputs can be linked, directly or indirectly, to identifiable individuals.

7.4.2 Purpose limitation

For zone-level and temporal operations on emotion vectors, only aggregate use of emotion is allowed, not writing analytics. No tracking, re-identification or profiling of any individual is done. On-device buffer size: Only last 10 frames (about 0.4 seconds) are stored. All The frames are thrown away once classified. No facial picture data is saved.

7.4.3 Transparency

Signage is posted in retail settings that have the system to let customers know about the system. An explanation of what emotion-aware analytics is, why it is being used and that no data is being retained and no one is being identified.

7.4.4 Lawful basis and consent

Any deployment must identify a lawful basis for processing before collection begins, and must determine whether consent is required in the applicable jurisdiction. Where consent is the basis relied upon, it must be freely given, specific, informed and withdrawable, and a practical opt-out must be available to customers who do not wish to be processed. Re-identification and inference risk Derived affective outputs may remain sensitive even in the absence of facial images. If emotion vectors are timestamped and associated with a camera position, they may be linkable to an individual through till records, loyalty identifiers, entry logs or CCTV retained for other purposes. Deployments should therefore assess linkage risk across all systems operating in the same premises, and not only within the emotion-analytics pipeline. Bias and fairness. The demographic composition of the evaluation data constrains what can be claimed about performance across populations, as noted in limitation L5. Differential error rates across demographic groups may produce unequal treatment if emotion outputs inform commercial decisions. A deployment should therefore evaluate performance by demographic group in the intended population before use, and should document the outcome. Governance and organizational controls. The technical measures described above address data minimization. They do not by themselves address retention schedules, access control, audit logging, vendor and processor obligations, breach notification, or the handling of data-subject requests. These require organizational controls that sit outside the architecture.

7.5 Causal Validity and Proxy Corruption

7.5.1 Proposed causal structure

The proposed structure is expressed as a directed acyclic graph in which emotional expression is a candidate cause of purchase, and four variables act as common causes of both: prior purchase intent, budget availability, social context, and product attractiveness. Each has a directed edge into emotional expression and a directed edge into purchase, and each therefore opens a back-door path between them. The causal estimand of interest would be the average change in purchase probability under an intervention on the emotional state, with the confounders above held fixed. Identifying that estimand would require, at minimum: that the confounding set is complete and measured; that the intervention is well defined; that positivity holds across the relevant covariate strata; and that the outcome is observed at the individual level. None of these conditions is satisfied by the present study, in which no purchase outcome is observed and no intervention is implemented. Causal validity is therefore not demonstrated here, and the framework below is a specification for future work rather than a result.

Four variables were found that were confounded with emotional expression and were in turn unique factors: values, moods, attitudes, and behaviors. Prior purchase intent of customers who enter with the goal of purchase may voice their intent to purchase. (1) positive emotions when they locate the desired product; (2) budget availability financial conditions impact. Eccentricity of the emotional state and of the purchasing independently; (3) social context shopping companion's influence. Routinely express their emotions and make purchases; (4) Repeat customers may have a routine or pattern for expressing their emotion and buying products; Buy neutral expressions that are from habit. The emotion purchase correlation needs to be validated with point-of-sale data.

data from transactions that are processed in a controlled experimental design in advance of retail operators acting on emotion-based data analytics.

7.5.2 Proxy Corruption Mechanisms

The following proxy-corruption mechanisms are advanced as hypotheses for future empirical investigation, and were not observed in the present study. Sustained optimization may have three effects that degrade the emotion purchase proxy identified: (1) environmental mood induction ambient music, light, scent creates positive mood. (2) expression modification (customers who know about); (3) emotions without purchase intent increase. (4) temporal decoupling as an example of suppressing or enhancing expressions, while decreasing the validity of the signal. The causal driver changes when stores optimize for emotion, and the emotion becomes an effect. If the store optimizes for emotion, then the causal driver changes, and the emotion becomes an effect purchase correlation decay. Continuous monitoring of emotion purchase correlation coefficient with automated. Alerts in case correlation is below a pre-defined threshold. Emotion analytics should be used as one.

7.6 Limitations

There are five limitations to this study which restrict the conclusions that can be drawn from it.

L1. Posed vs. spontaneous expressions: The CK+ consists of poses that are played under controlled circumstances. Spontaneous retail expressions are weak, unstable and have other temporal dynamics. The 92.42% benchmark accuracy should therefore not be read as an estimate of retail performance in either direction.

L2. Emotion taxonomy mismatch: The researcher, Ekman, has identified seven basic emotions that are expressed by everyone, including angry, disgust, fear, happy, sad, surprised, and disgusted. surprise, neutral) may not adequately represent the affective states that are most important in the context of retail, such as interest. The four factors that do not show up on the typical datasets are boredom, consideration, hesitation, and satisfaction.

L3. No transaction validation: No actual purchase data was used to validate the emotion purchase correlation. The shopping decision analysis module remains a theoretical framework requiring empirical validation.

L4. No retail pilot: The system was evaluated on benchmark datasets only. Real-world deployment in a retail environment was not conducted.

L5. Demographic bias: CK+ is predominantly composed of Caucasian subjects. Model performance on South Asian, East Asian, and African facial morphologies was not evaluated.

8. CONCLUSION

This paper made five contributions to emotion-aware retail analytics. Among the evaluated temporal architectures, the Attention BiLSTM achieved the highest observed accuracy of 92.42% on the CK+ temporal sequences under the experimental conditions considered, and exceeded the Deep BiLSTM by 6.06 percentage points while using fewer parameters. The comparison with the CK+ spatial baseline is not controlled and is not presented as a measure of the benefit of temporal modelling. The two-stage transfer learning approach successfully separated spatial and temporal learning between datasets. The distribution shift analysis identified that 15-35% accuracy can be expected in retail practice, setting performance expectations. The privacy preserving architecture showed that DPDP Act 2023 compliance can be achieved using on-device processing without loss of analytical power. The causal validity and proxy corruption analyses provided limits on the claims that can be made with emotion-based retail analytics. The following research will overcome these limitations by: (1) gathering consented retail expression datasets with spontaneous emotions; (2) creating retail-specific emotion taxonomies in line with retail marketing theory; (3) validating against point-of-sale data via controlled experimental design; (4) adapting to the retail training distribution via domain adaptation; (5) piloting with partner retailers to test performance and privacy-preserving architecture.

References

[1] Ekman P, Friesen WV. Constants Across Cultures in the Face and Emotion. J Pers Soc Psychol. 1971;17:124-129.

[2] Mehta D, Siddiqui MF, Javaid AY. Recognition of Emotion Intensities Using Machine Learning Algorithms: A Comparative Study. Sensors. 2019;19:1897.

[3] Bagane P, Vishal S, Raj R, Ganorkar T, Riya. Facial Emotion Detection Using Convolutional Neural Network. Int J Adv Comput Sci Appl. 2022;13.

[4] Zhao X, Shi X, Zhang S. Facial Expression Recognition via Deep Learning. IETE Tech Rev. 2015;32:347-355.

[5] Mellouk W, Handouzi W. Facial Emotion Recognition Using Deep Learning: Review and Insights. Procedia Comput Sci. 2020;175:689-694.

[6] Fatima SA, Kumar A, Raoof SS. Real Time Emotion Detection of Humans Using Mini-Xception Algorithm. IOP Conf Ser Mater Sci Eng. 2021;1042:012027.

[7] Giannopoulos P, Perikos I, Hatzilygeroudis I. Deep Learning Approaches for Facial Emotion Recognition: A Case Study on Fer-2013. Hatzilygeroudis I, Palade V, editors . Advances in hybridization of intelligent methods: Models, systems and applications. Cham: Springer International Publishing. 2017;85:1-16.

[8] Jaiswal A, Raju AK, Deb S. Facial Emotion Detection Using Deep Learning. In: Proceedings of the INCET. In 2020 International Conference for Emerging Technology (INCET). IEEE. 2020:1-5.

[9] Khare SK, Bajaj V. Time-Frequency Representation and Convolutional Neural Network-Based Emotion Recognition. IEEE Trans Neural Netw Learn Syst. 2021;32:2901-2909.

[10] Talegaonkar I, Joshi K, Valunj S, Kohok R, Kulkarni A. Real Time Facial Expression Recognition Using Deep Learning. In Proceedings of the International Conference on Communication and Information Processing (ICCIP). 2019. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3421486

[11] Siam AI, Soliman NF, Algarni AD, Abd El-Samie FE, Sedik A. Deploying Machine Learning Techniques for Human Emotion Detection. Comput Intell Neurosci. 2022;2022:8032673.

[12] Mehendale N. Facial Emotion Recognition Using Convolutional Neural Networks (FERC). SN Appl Sci. 2020;2:446.

[13] Alhalaseh R, Alasasfeh S. Machine-Learning-Based Emotion Recognition System Using EEG Signals. Computers. 2020;9:95.

[14] Badrulhisham NA, Mangshor NN. Emotion Recognition Using Convolutional Neural Network (CNN). J Phys Conf Ser. 2021;1962:012040.

[15] Chakriswaran P, Vincent DR, Srinivasan K, Sharma V, Chang CY, et al. Emotion AI-Driven Sentiment Analysis: A Survey, Future Research Directions, and Open Issues. Appl Sci. 2019;9:5462.

[16] Kabra B, Nagar C. Attention-Emotion-Embedding BiLSTM-GRU Network-Based Sentiment Analysis. J Integr Sci Technol. 2023;11:563.

[17] Donovan RJ, Rossiter JR, Marcoolyn G, Nesdale A. Store Atmosphere and Purchasing Behavior. J Retail. 1994;70:283-294.