An Intelligent Surveillance System Using Deep Facial Expression Recognition

Livhuwani Mutshafa 17010588

Supervisor: Dr. B Moyo

Submitted in partial fulfillment of the requirements for the degree of Master of Science in eScience

in the

Department of Mathematical and Computational Sciences, Faculty of Science, Engineering, and Agriculture

University of Venda, Thohoyandou

27 February 2026

Declaration

I, Livhuwani Mutshafa 17010588, affirm that this research is entirely my original work. It is being presented for the degree of Master of Science in eScience at the University of Venda, Thohoyandou. This research has not been previously submitted for any degree or examination at any other institution.

mutshafa .l.

Livhuwani Mutshafa

17010588

27 February 2026

3 March 2026

Dr. B Moyo

Abstract

Surveillance systems are critical tools for maintaining security, enhancing public safety, and safeguarding assets in diverse settings, from public spaces to private facilities. Despite their importance, these systems often face challenges that require human oversight. Recent studies have explored deep learning techniques to address such challenges, primarily focusing on face recognition and anomaly detection in static images. This study proposes a deep learning approach for detecting and interpreting facial expressions in dynamic images to enhance surveillance applications. The methodology involved a comprehensive literature review, dataset preprocessing, development of deep learning models, and rigorous model evaluation. A fine-tuned MobileNetV2 and a hybrid MobileNetV2–LSTM models were designed to capture both spatial and temporal features of facial expressions. The models were trained on benchmark datasets, including the Amsterdam Dynamic Facial Expression Set (ADFES) and the Chinese Face Dataset with Dynamic Expressions, and evaluated using accuracy, precision, recall, and F1-score metrics. Results demonstrated that the MobileNetV2–LSTM model significantly outperformed the standard MobileNetV2, achieving 95% accuracy, 95% precision, 95% recall, and 95% F1-score, highlighting the advantages of temporal modeling. The models maintained high computational efficiency, achieving 43.09 frames per second and a per-frame inference time of 0.0232 seconds, indicating strong real-time feasibility. This study contributes to intelligent surveillance by providing a highly reliable facial expression recognition framework for dynamic scenarios, with future work focusing on real-time deployment, expanded datasets with diverse ethnicities, and enhanced robustness under challenging surveillance conditions.

Keywords: ADFES, Chinese dataset, Deep learning, Dynamic images, Surveillance systems

Acknowledgements

I am truly grateful to the Almighty for providing me with the strength and determination to complete this research. I extend my heartfelt thanks to the National e-Science Postgraduate Teaching and Training Platform (NEPTTP) for their sponsorship. Additionally, I wish to express my appreciation to my supervisor, Dr. B. Moyo, for his invaluable guidance and support throughout the research. Lastly, my supportive family and friends for their unwavering encouragement.

Table of Contents

List of Figures

List of Abbreviations

AFDES	Amsterdam Dynamic Facial Expression Set
LSTM	Long Short-Term Memory
LBP	Local Binary Patterns
HOG	Histogram of Oriented Gradients
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
SVM	Support Vector Machine
k-NN	k-Nearest Neighbors
PCA	Principal Component Analysis
GPU	Graphics Processing Unit
HPC	High performance computing
DL	Deep Learning
FPS	frames per second
IPS	Images per second
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
FER	Facial Expression Recognition

Research Outputs

The following list provides a record of the research outputs associated with this study. It includes work that has been presented at a symposium and accepted for conference presentations.

1. Part of this research was presented at the *Univen-UL Symposium on Regional Promotion of Mathematical and Computational Sciences*, held on 23–24 June 2025, at the Research Conference Centre, University of Venda, South Africa. The presentation, titled **"An Intelligent Surveillance System Using Deep Facial Expression Recognition,"** focused on preliminary results that informed the development of this mini-dissertation.

2. Part of this research was presented at the IFIP-UNIVEN-CSIR *International Cybersecurity Conference 2025*, held on 11–12 December 2025 at CSIR ICC, Pretoria, South Africa, and published in the conference proceedings. The paper, titled **"Towards Facial Expression Analysis for Enhanced Threat Detection in Surveillance,"** incorporates core findings from this study with expanded analysis. DOI:https://doi.org/10.1007/978-3-032-13075-4_4

Chapter 1 Introduction

1.1 Background of the Study

Advances in technology and data analytics have led to significant progress in surveillance systems [12]. The primary function of these systems is to monitor environments and capture activities within a defined period [70]. However, their full potential remains underutilised because they rely heavily on human intervention for real-time threat detection and response [68]. This dependence on human operators poses major challenges, particularly when immediate action is required to mitigate threats such as theft, kidnapping or murder. To enhance effectiveness, surveillance systems must be capable of responding autonomously in the absence of human input [45].

The dynamic nature of surveillance environments presents unique challenges for facial expression recognition systems. These include variations in lighting, camera angles, partial facial occlusions and rapid changes in expressions. Such conditions differ greatly from the controlled laboratory environments typically used to develop and test most facial expression systems, where lighting remains constant and subjects face the camera directly [21, 69, 74].

Despite these challenges, facial expression recognition remains a crucial component of intelligent surveillance. In public institutions such as schools, airports and transportation hubs, security personnel are required to monitor thousands of faces simultaneously [2]. In such contexts, automated systems can detect subtle emotional and behavioural cues that human operators may overlook. These cues may indicate potential threats such as suspicious activity, stress or aggression [62]. An effective recognition system in dynamic environments can therefore enable real-time responses to security risks, enhance public safety and reduce the cognitive load on human operators who struggle with continuous monitoring of multiple video feeds [65]. Addressing these challenges and improving the robustness of recognition systems are essential steps towards advancing modern surveillance capabilities [62].

Earlier studies primarily employed traditional machine learning methods, where facial

expression features were manually extracted and classified using separate algorithms [15]. More recently, deep learning techniques such as Convolutional Neural Networks (CNNs) have shown considerable improvement, as CNNs automatically learn features from raw data and use them as input for emotion recognition tasks [15].

According to Ekman et al. [18], human emotions can be categorised into six basic types: sadness, happiness, fear, surprise, anger and disgust. They further explained how specific facial features such as the lips, eyes, eyebrows and eyelids change with each expression. Understanding facial expressions, however, requires examining how these features dynamically shift with emotions. For instance, the position and movement of the lips may convey sadness or happiness, while the eyes reveal subtle changes in gaze direction. Movements of the eyelids and pupil dilation also reflect underlying emotional states.

Existing studies have made notable progress in facial recognition and anomaly detection within surveillance systems. For example, Singh et al. [58] demonstrated the effectiveness of facial recognition systems in surveillance applications, while Choudhry et al. [15] focused on improving anomaly detection. However, these studies relied on models trained on static images, which fail to capture emotional variations in dynamic environments typical of surveillance systems. Although static image-based models perform well in controlled conditions, they often struggle to adapt to real-world scenarios where emotions are continuously changing [33].

Therefore, this study aims to develop an intelligent surveillance system using a deep CNN model trained on dynamic images. By applying deep learning techniques to analyse dynamic video streams, the proposed system seeks to overcome the limitations of static image-based models and improve accuracy in capturing changes in facial expressions within real-time surveillance contexts.

1.2 Problem Statement

Surveillance systems rely heavily on humans to monitor and detect potential security threats in real time. However, this approach is often inefficient, prone to human error and may result in delayed responses, allowing threatening acts to occur before they can be prevented. There is therefore a critical need to enhance the capabilities of surveillance

systems by incorporating intelligent techniques that can automatically detect and respond to suspicious behaviour promptly, without relying solely on human intervention.

According to the literature, facial recognition techniques and deep learning model can be leveraged to analyse facial expressions captured by surveillance systems to address this challenge [26, 80]. Facial expressions provide valuable indicators of an individual's emotional state, intentions and potential for threatening behaviour. Developing an intelligent surveillance system that accurately extracts relevant facial expression features and integrates them with a deep learning model can strengthen threat detection capabilities. This approach may enable real-time responses to potential security threats without the need for constant human monitoring.

This study therefore proposes the use of deep learning techniques to enhance surveillance systems through real-time facial expression monitoring. The use of deep learning will enable automated analysis of facial expressions, allowing for timely interventions and reducing dependence on manual human involvement in surveillance operations.

1.3 Research Questions

The research is based on the following questions:

i. How can a deep learning model for facial expression recognition be developed to achieve high accuracy in detecting and classifying a wide range of facial expressions in dynamic environments?
ii. How can a deep learning model for facial expression recognition be designed to achieve robust, real-time performance in dynamic surveillance environments?
iii. How does the intelligent surveillance system developed using deep facial expression recognition compare to existing surveillance systems in terms of accuracy, speed and reliability when evaluated on benchmark datasets?

1.4 Research Aim and Objectives

This section outlines the aims and objectives, which will provide directions for addressing the research questions.

1.4.1 Research Aim

This study aims to develop an intelligent surveillance system that utilizes deep facial expression recognition techniques, focusing on accurately detecting and interpreting emotions in real-time for enhanced security applications, while considering ethical and privacy implications.

1.4.2 Objectives

The objectives of the study are:

i. To investigate the development of intelligent surveillance systems using facial expression and deep learning techniques.
ii. To investigate and optimize the performance of the deep learning model by evaluating different network architectures and training strategies to improve the accuracy and robustness of facial expression recognition in dynamic surveillance environments.
iii. To identify potential ethical concerns, including privacy threats and demographic bias, in the development of the facial expression model, and establish guidelines for responsible use.
iv. To evaluate the performance and effectiveness of the intelligent surveillance system using benchmark datasets, through conducting extensive experiments and benchmarking against existing surveillance systems.

1.5 Scope of the Study

This study employs deep learning approaches to develop a surveillance system capable of recognising human facial expressions. It aims to interpret facial expressions in dynamic images, enabling surveillance systems to understand expressions despite temporary variations. The study moves beyond static image analysis towards a more realistic approach to facial expression recognition. In addition, it addresses the challenge of occlusion in facial expression recognition.

The research covers data preprocessing, model selection, model training, hyperparameter tuning, evaluation and validation using publicly available datasets.

Although the system will not be deployed in a real-world surveillance environment, the study seeks to contribute to existing literature and provide a foundation for future surveillance applications in visually challenging and dynamic contexts.

1.6 Significance of the Study

Surveillance systems are essential components of modern security, helping to monitor environments and capture real-time activities [70]. They also support investigations and evidence verification by allowing past events to be retrieved when required. However, despite their potential, many surveillance systems remain underutilised. Detecting crimes such as theft, kidnapping or assault often depends on human operators actively monitoring surveillance footage, which is not always practical [45]. In some cases, operators may respond promptly to threats, but in others, incidents are only identified after they have occurred.

This study aims to address these challenges by integrating deep learning-based facial expression recognition into surveillance systems. By detecting emotions such as distress, fear and aggression in real time, the system can automatically flag unusual behaviour, prompt faster responses and reducing the need for constant human monitoring [65]. Unlike traditional models that analyse static images, this research focuses on dynamic images, making it more effective for real-world surveillance situations where facial expressions change rapidly.

In addition to contributing to the growing body of research on deep learning for facial expression recognition, this study has significant implications for public safety. Enhancing crime prevention and improving early threat detection can help create safer environments and reduce the likelihood of incidents occurring.

1.7 Structure of the Mini-Dissertation

The mini dissertation consists of five chapters.

- **Chapter 1: Introduction**

  This chapter presents the background of the study, the problem statement, aim

and objectives, significance of the study, scope, and elaboration of ethical considerations.

- **Chapter 2: Literature Review**

  The review focuses on existing research on facial expression recognition and identifies research gaps.

- **Chapter 3: Research Methodology**

  Describes the dataset, preprocessing methods, model development, hyperparameter tuning, evaluation metrics, and ethical considerations.

- **Chapter 4: Results and Discussion**

  Present experimental results, evaluate model performance and compare findings with existing studies.

- **Chapter 5: Conclusion and Recommendations**

  Summarizes key findings, discusses limitations, and offers suggestions for future work.

1.8 Conclusion

This chapter has presented the introduction and background of the study. It outlined the problem statement, research aim and objectives, research methodology, significance, scope, and ethical considerations. The organisation of the mini-dissertation was also discussed. The next chapter provides a review of the literature relevant to this stud

Chapter 2 Literature Review

2.1 Facial Expressions and Facial Expression Recognition

Facial expressions are observable manifestations of internal emotional states, produced through the coordinated activation of facial muscles. They represent a primary channel of non-verbal communication, conveying essential information about an individual's emotional state, intentions and social interactions [18, 50]. According to Ekman et al. [18], facial expressions can be broadly categorised into six fundamental emotions: happiness, sadness, fear, anger, disgust and surprise. These emotions are expressed through dynamic variations in facial characteristics, including lip movement and positioning, eye motion, eyebrow activity and eyelid behaviour.

Despite their universality in human interaction, the theoretical foundations of facial expressions remain widely debated among scholars. Ekman's universality hypothesis proposes that basic emotions are expressed through consistent facial configurations across cultures [18], forming the basis for much of contemporary research in facial expression recognition. However, this view has been critically challenged by constructionist theorists, who argue that facial expressions are culturally and contextually influenced rather than biologically predetermined [10].

Several complexities further complicate the analysis and interpretation of facial expressions. Ambiguity arises when identical facial expressions are interpreted as different emotions depending on contextual factors [51]. Individual variability also affects recognition accuracy, as the intensity and manifestation of expressions differ significantly across demographic groups and individuals [38]. The distinction between spontaneous and posed expressions presents another challenge, as genuine emotional displays differ systematically from deliberate ones [49]. In addition, temporal dynamics involving the onset, apex and offset phases of an expression provide critical diagnostic information that must be accurately modelled for effective interpretation [73]. Collectively, these factors highlight the multifaceted and context-dependent nature of facial behaviour and underscore the need for advanced computational models capable of capturing these subtle and temporally dynamic variations.

Facial Expression Recognition (FER) refers to the automated process of identifying human emotions through the analysis of facial movements and visual features [38]. A typical FER system consists of three primary stages: face detection, feature extraction and emotion classification [38, 52, 80]. Over time, the field has evolved from traditional machine learning approaches, which relied on handcrafted features, to advanced deep learning architectures capable of automatically learning discriminative representations from raw image data. These technological advances have significantly improved the accuracy, robustness and generalisability of FER systems, enabling their application in various fields, including human–computer interaction, psychological assessment and intelligent surveillance [14, 31].

According to Ekundayo et al. [19], integrating facial expression recognition into surveillance systems offers a promising approach to enhancing security measures. Traditional surveillance systems rely heavily on human operators for threat detection, which introduces challenges related to operator fatigue, subjective interpretation and delayed response times [70]. Automated facial expression recognition enables surveillance systems to identify subtle emotional indicators that may suggest suspicious activity, stress or violent behaviour before they escalate into security threats [38].

This chapter provides an overview of techniques previously employed in facial expression recognition, tracing the evolution of the field and examining its theoretical and practical implications, particularly within the context of surveillance applications.

2.2 Traditional Machine Learning Approaches to Facial Expression Recognition

Facial expression recognition primarily relied on traditional machine learning approaches before the widespread adoption of deep learning techniques [14]. These approaches are generally categorised into three types: appearance-based, hybrid and geometric-based methods. They typically involve two main steps, feature extraction using handcrafted techniques, followed by classification using traditional machine learning algorithms [14]. Common handcrafted methods for capturing texture and facial

appearance include Gabor wavelets, Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP).

In earlier studies, Shan et al. conducted a comprehensive investigation using Local Binary Patterns for facial expression recognition, demonstrating the method's robustness under varying lighting conditions [56]. However, their work struggled with head-pose variations and failed to capture subtle expressions. Similarly, Lades et al. explored the use of Gabor wavelets, where Gabor filters mimic the behaviour of cells in the human visual cortex by capturing spatial frequency and orientation information [34]. Despite their effectiveness, Gabor-based approaches proved computationally expensive.

Dalal and Triggs introduced the use of Histograms of Oriented Gradients for pedestrian detection [17], which was later adapted by Happy and Routray for facial expression recognition to capture edge orientations. After extracting features, Support Vector Machines (SVMs) were used for classification. Their findings showed that SVM classifiers, particularly those using Radial Basis Function kernels with features derived from salient facial patches, achieved high recognition accuracy while maintaining computational efficiency and robustness across datasets and image resolutions.

Bartlett et al. further investigated the use of SVMs for classifying facial expressions from video data. The authors trained SVM classifiers on features extracted from facial images, achieving a high accuracy rate of 96%. Their results showed that SVMs were computationally efficient, especially when combined with feature selection methods [11]. Similarly, Nazir et al. employed SVMs in conjunction with HOG features to classify facial expressions in static images [42]. Their results indicated that SVMs perform well when paired with carefully engineered features. The key strength of SVMs lies in their ability to use kernel functions that project features into higher-dimensional spaces, enabling the separation of non-linearly separable data. However, SVMs perform poorly with very large datasets and are unable to capture temporal dependencies effectively.

Tree-based methods have also been used in facial expression recognition because of their computational efficiency and interpretability. Salmam et al. applied a decision tree on the CK+ and JAFFE datasets using geometric features and achieved competitive results through ensemble voting strategies [53]. Random forests have also shown promising results when combined with traditional feature extraction methods. For example, Zhao

et al. combined random forests with Local Binary Patterns for emotion classification [48]. Despite these results, tree-based methods have limitations when dealing with complex hierarchical features, subtle temporal changes and high-dimensional feature spaces in facial expression recognition.

Sohail and Bhattacharya presented one of the earliest traditional approaches for classifying six basic facial expressions from frontal face images using a K-Nearest Neighbour (KNN) classifier [61]. The algorithm classifies expressions based on the majority class of the k closest training samples in the feature space, making it particularly suitable for small datasets with limited training data. However, KNN suffers from sensitivity to irrelevant features and high computational demands, which limit its applicability in real-time systems [3].

Although traditional approaches demonstrated satisfactory performance, they exhibited significant limitations when dealing with pose variations, real-world lighting conditions and partial occlusions [14]. Moreover, SVMs faced scalability challenges with large datasets and were unable to model temporal dependencies effectively. These constraints stemmed from the reliance on handcrafted features and manual feature engineering. Consequently, these limitations motivated a paradigm shift towards deep learning-based methods, which automatically learn hierarchical feature representations directly from raw data and are inherently capable of capturing both spatial complexity and temporal dynamics in facial expressions [31].

2.3 Deep Learning Approaches to Facial Expression Recognition

Deep learning is an advanced subset of machine learning inspired by the structure and function of the human brain, particularly artificial neural networks [4, 66]. Unlike traditional methods that rely on handcrafted features, deep learning models automatically learn hierarchical feature representations from raw data through multiple layers of abstraction [4]. This capability allows them to capture complex spatial and temporal dependencies directly from images or video sequences, making them particularly effective for visual recognition tasks. Deep learning emerged to address the

limitations of conventional algorithms, which often struggled to generalize under variations in pose, illumination, and occlusion [14]. The widespread availability of large-scale annotated datasets, coupled with advances in computational power, especially Graphics Processing Units (GPUs), has further accelerated its adoption in computer vision applications. As a result, deep learning architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks have achieved state-of-the-art performance in facial expression recognition tasks [31, 35].

2.3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) have emerged as powerful tools for facial expression recognition due to their ability to automatically learn hierarchical representations directly from raw image data. Several CNN architectures have been proposed specifically for facial expression recognition tasks [31].

A study by Yolcu et al. [75] introduced a CNN architecture consisting of four networks, which they applied to the RAFD dataset. Their research aimed to improve facial expression recognition accuracy, and their approach achieved state-of-the-art performance in the facial expression recognition domain. By leveraging the RAFD dataset, which contains diverse facial expressions representing various emotions, Yolcu et al. [75] demonstrated the effectiveness of their CNN architecture in accurately recognizing facial expressions. Their findings highlighted the potential of deep learning techniques to advance the field of facial recognition and contribute to the development of more robust and reliable systems for real-world applications.

Similarly, Hussain et al. [26] developed a real-time facial emotion classification system using a CNN based on the VGG-16 architecture. The model was trained and evaluated on the Karolinska Directed Emotional Faces (KDEF) dataset and achieved an accuracy of 88% in detecting and classifying facial expressions. Other CNN-based studies have demonstrated strong performance in controlled environments, confirming the effectiveness of deep convolutional architectures for feature extraction and emotion classification.

Despite these successes, several limitations of CNN-based approaches have been

identified. These models typically require large volumes of data to perform optimally, making data collection and labeling a significant challenge [30, 54, 57]. Furthermore, CNNs are often regarded as black-box systems, offering limited interpretability regarding the internal decision-making process [1]. Models trained on constrained laboratory datasets, such as KDEF, frequently struggle to generalize to real-world conditions involving diverse poses, lighting variations, and occlusions [35, 31]. Additionally, architectures such as VGG-16 impose high computational and memory demands, which may restrict deployment on embedded systems [68].

2.3.2 Object Detection-Based Approaches

Object detection-based approaches for facial expression recognition leverage advancements in object detection frameworks to first locate faces within an image and then perform expression classification. These methods are particularly valuable in surveillance contexts, where multiple faces may be present in varying orientations and scales.

Algorithms such as You Only Look Once (YOLO) have proven effective in detecting key facial features. Venkateshwarlu et al. [26] demonstrated that integrating YOLO with CNNs can significantly enhance the performance of facial expression recognition systems. Their approach achieved faster processing times compared to traditional two-stage methods while maintaining high accuracy in challenging environments.

Face detection serves as a crucial initial step for any facial recognition task, providing the foundation by accurately identifying and localizing faces within images or video frames [46]. Once the face is detected, facial landmarks such as the eyes, lips, mouth, and nose can be identified, which play a critical role in facial expression recognition. Without precise face detection, the interpretation and understanding of emotions may be compromised. Therefore, the effectiveness of face detection techniques directly influences the overall accuracy of emotion recognition systems [46].

2.3.3 Temporal Modeling Approaches

Temporal modeling approaches recognize that facial expressions are dynamic and evolve, particularly in video sequences. These methods aim to capture the temporal dynamics of expressions to improve recognition accuracy.

Jeong et al. [29] proposed the Deep Joint Spatio-Temporal Network (DJSTN) to enhance both accuracy and efficiency in emotion detection from facial expressions. The model employs 3D convolutional layers to simultaneously extract spatial and temporal features from video sequences, addressing the limitations of earlier methods that process these features separately. DJSTN integrates 23 dominant facial landmarks to represent key facial muscle movements, which are fused via a joint fusion classifier to improve recognition performance. Evaluated on three benchmark datasets, CK+, MMI, and FERA, the model achieved 99.21% recognition accuracy on CK+, demonstrating the effectiveness of combining spatio-temporal and geometric features. However, this high performance may be influenced by the controlled nature of CK+, raising concerns about generalization to real-world, unconstrained scenarios.

Similarly, Yin et al. [74] developed a dynamic emotion recognition system using a custom CNN-RNN architecture applied to the ADFES dataset. The model achieved 58% validation accuracy and 44% accuracy on unseen test data, highlighting challenges in generalization. They introduced PN accuracy, a metric for distinguishing positive and negative emotions, which reached 76% on validation data, indicating that their system could effectively capture emotional valence. Their temporal visualization approach provided insights into how emotions transition from neutral expressions over time. However, the model struggled with negative emotions, particularly contempt, and exhibited limited depth in feature extraction compared to larger pre-trained networks such as ResNet or ImageNet based architectures. Additionally, the system was not optimized for real-time inference, limiting its applicability in surveillance and other resource-constrained environments.

Vosta et al. [70] demonstrated a hybrid CNN-convLSTM approach, where features extracted by a ResNet50 CNN from video frames were passed to a convLSTM network to capture temporal dependencies. Their model achieved 81.71% accuracy, illustrating the effectiveness of combining spatial and temporal modeling techniques for dynamic facial expression recognition.

Despite these advances, temporal modeling and object detection-based approaches have notable limitations. Performance is highly dependent on the accuracy of initial face detection, with errors propagating through the system [5]. Heavily occluded faces or crowded scenes can reduce detection reliability, while multi-stage pipelines introduce

higher computational overhead and latency [71, 78]. These approaches also require extensive annotated datasets, which are costly and labor-intensive to produce [78]. Variations in face orientation, scale, and illumination further challenge robustness, limiting system reliability in unconstrained, real-world environments [78].

2.4 Transfer Learning Approaches

Transfer learning is a machine learning technique in which knowledge acquired from solving one task is leveraged to improve performance on a different but related task [44]. This approach exploits the hierarchical nature of deep neural networks, where lower layers learn general features such as edges and textures, while higher layers capture task-specific patterns [44]. Transfer learning addresses three key challenges: limited availability of labeled training data in specialized domains, the substantial computational cost of training deep networks from scratch, and the need for models to generalize across different data distributions [32]. In practice, transfer learning operates through two primary mechanisms: feature extraction, where pre-trained weights are frozen and used as fixed feature extractors, and fine-tuning, where pre-trained models are adapted to new tasks through continued training [32, 44].

Several studies have demonstrated the effectiveness of transfer learning for facial expression recognition. Mollahosseini et al. fine-tuned CNNs pretrained on ImageNet for emotion classification datasets such as CK+, JAFFE, and AffectNet, showing improved performance compared to training from scratch, particularly with small datasets [40]. Similarly, Minaee et al. employed ResNet and SqueezeNet architectures to achieve state-of-the-art performance on the FER2013 dataset by reusing low-level visual features such as edges, contours, and textures [39].

Comparative studies have examined which architectures are most suitable for facial expression recognition. Li et al. [36] compared VGG16, InceptionV3, ResNet50, and DenseNet-121 on the RAF-DB and AffectNet datasets. ResNet50 offered a good balance between accuracy and computational efficiency, with intermediate layers extracting meaningful spatial features. This supports the transfer learning principle that pretrained networks can extract transferable spatial representations [32]. However, the computational demands of architectures such as ResNet50 and VGG16 may limit their

deployment in resource-constrained environments.

Hybrid models that combine spatial and temporal modeling have also been explored. Zhang et al. [76] proposed a CNN-LSTM model pretrained on ResNet18, where spatial features extracted by ResNet18 were passed to an LSTM to capture temporal dependencies in facial expression sequences. The model performed strongly on the CK+ dataset, capturing transitions between expressions effectively, but its computational intensity limits real-time applications. Bargshady et al. [9] developed an EJH-CNN-BiLSTM model using a fine-tuned VGG-Face pre-trainer and Principal Component Analysis (PCA) to enhance computational efficiency. Their approach demonstrated potential for real-time implementation in online systems, but it may not fully address the challenges of surveillance systems such as large-scale video streams, multiple subjects, occlusions, and uncontrolled lighting.

Transfer learning has consistently shown improved performance compared to training from scratch. Oztel et al. [43] evaluated VGG16 on the RaFD dataset with and without transfer learning, confirming that the pretrained model performed better. Gondkar et al. [22] applied multiple pretrained architectures, including Xception, ResNet, and VGG, on the CK+ dataset and observed higher training and validation accuracy for ResNet-based models. Kosti et al. highlighted the advantages of combining transfer learning with data augmentation and domain adaptation to reduce dataset bias and improve cross-dataset generalization.

MobileNet architectures have been designed for computational efficiency. Howard et al. developed MobileNet using separable convolutions [24], while Huang et al. implemented MobileNetV2 combined with LSTMs using transfer learning and fine-tuning. This approach achieved competitive accuracy on real-time facial expression recognition tasks while maintaining minimal inference time, making it suitable for mobile and embedded systems [25].

Despite its advantages, transfer learning has limitations. Performance may degrade when source and target domains are dissimilar, and overfitting remains a concern when fine-tuning on very small datasets. Nevertheless, transfer learning provides a balance between high accuracy and computational efficiency [39, 25, 40].

Feature extraction is a fundamental step in any recognition task, aiming to identify patterns or characteristics in data that are relevant to the task. Saadi et al. [52] categorize feature extraction into handcrafted and learned features. Handcrafted features rely on manually designed algorithms and domain expertise to identify patterns such as edges, textures, and geometric shapes. These features are interpretable and computationally efficient but often struggle to capture subtle patterns and require substantial human effort. Learned features, on the other hand, are automatically discovered by deep learning models, progressively extracting information from simple edges to high-level semantic patterns [52]. In facial expression recognition, learned features have consistently outperformed handcrafted ones, particularly in real-world scenarios where lighting, pose, and occlusion vary, because they adaptively learn the most informative and discriminative representations directly from data [52].

2.5 Applications in Surveillance Systems

Deep learning techniques have been increasingly applied to surveillance tasks such as object recognition, crowd analysis, and action recognition. Sreenu and Saleem Durai [62] conducted a survey comparing various deep learning methods and proposed a model called AMDN to detect anomalies in video sequences. The model utilized autoencoders to learn features and identify abnormal patterns. The authors noted that AMDN requires additional computational resources for video analysis in surveillance systems. They also emphasized the significance of deep learning approaches, including Convolutional Neural Networks (CNNs), in addressing the challenges of analyzing crowded scenes in surveillance videos [62].

Singh et al. [58] developed a real-time method for detecting and identifying individuals in surveillance streams using CNNs and face recognition. Their system employed the VGGFace architecture with transfer learning, retraining the model on a custom dataset. The model achieved an accuracy of 78.54% in correctly identifying individuals. This study demonstrates the practical application of deep learning techniques in real-world surveillance scenarios, though its primary focus was on identity recognition rather than facial expression analysis.

Similarly, Singh et al. [11] presented a method for recognizing and identifying people in

live surveillance streams, also utilizing CNNs and face recognition. The deep learning neural architecture was combined with transfer learning, and the system achieved the same accuracy of 78.54% in identifying individuals.

Al-Amoudi et al. [6] developed an automatic attendance system based on face recognition, using FaceNet and MTCNN techniques. The system automated attendance recording for lecturers and achieved 87% accuracy in recognizing faces. While this study focused on attendance management rather than surveillance, it illustrates the potential of facial recognition technologies to handle multiple subjects in real-world applications.

Overall, these studies highlight the versatility and practical benefits of deep learning in surveillance contexts, demonstrating its ability to handle identity recognition, anomaly detection, and automated monitoring in complex environments. Limitations remain, including the computational demands of video analysis and the focus on identity rather than dynamic facial expression recognition.

2.6 Datasets in Facial Expression Recognition

According to Kopalidis et al. [33], the performance and generalizability of facial expression recognition models depend heavily on the quality of the training and evaluation datasets. A variety of datasets have been created and used in the literature, each with specific features suited to different aspects of facial expression recognition.

2.6.1 Static Image Datasets

Static image datasets consist of individual images that capture facial expressions at a specific moment. These datasets have been widely used in the development of facial expression recognition systems:

- **FER2013**: Contains 35,887 grayscale images of facial expressions categorized into seven emotion classes. The images were collected from the web under varying lighting conditions, occlusions, and head poses.
- **CK**: The Cohn-Kanade dataset includes sequences of posed facial expressions from 100 subjects, showing the progression from neutral to maximum expression.

- **JAFFE**: The Japanese Female Facial Expression dataset contains 213 images of seven expressions performed by ten Japanese female models. This dataset has been used as a benchmark in numerous studies.

- **RAFD**: The Radboud Faces Database consists of 67 high-resolution models displaying eight emotional expressions. It includes both frontal and profile views, making it suitable for multi-angle facial expression recognition.

2.6.2 Dynamic Image Datasets

Dynamic image datasets capture the temporal evolution of facial expressions, typically in the form of video sequences:

- **ADFES**: The Amsterdam Dynamic Facial Expression Set contains video recordings of 22 models displaying nine emotional expressions. This dataset is particularly valuable for studying temporal dynamics in facial expressions.

- **Chinese Face Dataset with Dynamic Expressions**: This dataset features video recordings of 60 men and 60 women, all aged between 18 and 28, capturing a variety of dynamic expressions.

- **CK+**: The Extended Cohn-Kanade dataset includes 593 sequences from 123 subjects, showing the progression from neutral to maximum expression. It is widely used for testing facial expression recognition models.

2.7 Research Gaps

The literature review reveals several gaps in facial expression recognition, particularly in real-world surveillance environments. Despite significant progress in static facial expression recognition, dynamic expression recognition in surveillance settings remains largely unexplored. Studies such as those by Singh et al. and Hussain et al. focused on static images in their models, neglecting the temporal evolution of facial expressions. However, in surveillance systems, capturing changes in facial expression over time is crucial for understanding emotional dynamics and identifying potential security threats.

Although facial recognition technologies have been incorporated into face verification systems, the potential of facial expression recognition for security purposes represents a key gap in the literature. Real-time processing, multi-subject tracking, and context-based emotion understanding in crowded and dynamic environments remain challenging areas reuiring further investigation.

Current temporal modeling approaches, such as those proposed by Jeong et al. [13] and Yin et al. [29], can capture dynamic facial expressions but are often computationally intensive. This restricts their applicability in real-time surveillance systems, where efficiency is critical. Developing lightweight, high-performance temporal models capable of accurately tracking expression dynamics without excessive computational demands remains an open research problem.

Overall, the literature shows that deep learning-based facial expression recognition techniques, particularly CNNs, perform well in detecting and classifying expressions in real-time video. However, most studies have focused on static images rather than dynamic sequences, limiting their effectiveness in video-based surveillance applications.

This study seeks to address these limitations by developing an intelligent surveillance system that leverages deep facial expression recognition technology to analyze dynamic video streams.

2.8 Conclusion

In conclusion, this chapter demonstrates the progression from traditional machine learning approaches to deep learning-based methods for facial expression recognition. Traditional approaches rely heavily on manual feature extraction, while deep learning methods, especially CNNs and transfer learning, have significantly improved recognition accuracy.

Despite these advances, most existing models depend on static images and are computationally intensive, which limits their effectiveness in real-time applications. To overcome these challenges, this study proposes a methodology that uses lightweight models and incorporates temporal dynamics, which are essential for video-based facial expression recognition.

The next chapter presents the methodology designed to address the research gaps identified in this chapter.

Chapter 3 Research Methodology

This chapter presents the research methodology employed to achieve the objectives outlined in Chapter 1. The study adopts a quantitative approach, leveraging deep learning techniques to address challenges in facial expression recognition. The methodology covers dataset acquisition and preprocessing, model selection and design, experimental setup and hyperparameter tuning, training, validation, testing procedures, and model performance evaluation.

To achieve the research objectives, this study employed two model architectures. The standard MobileNetV2 was used to establish a lightweight baseline, while a MobileNetV2–LSTM hybrid architecture was implemented to capture temporal dynamics. This dual approach aims to improve facial expression recognition accuracy while maintaining computational efficiency suitable for resource-constrained environments.

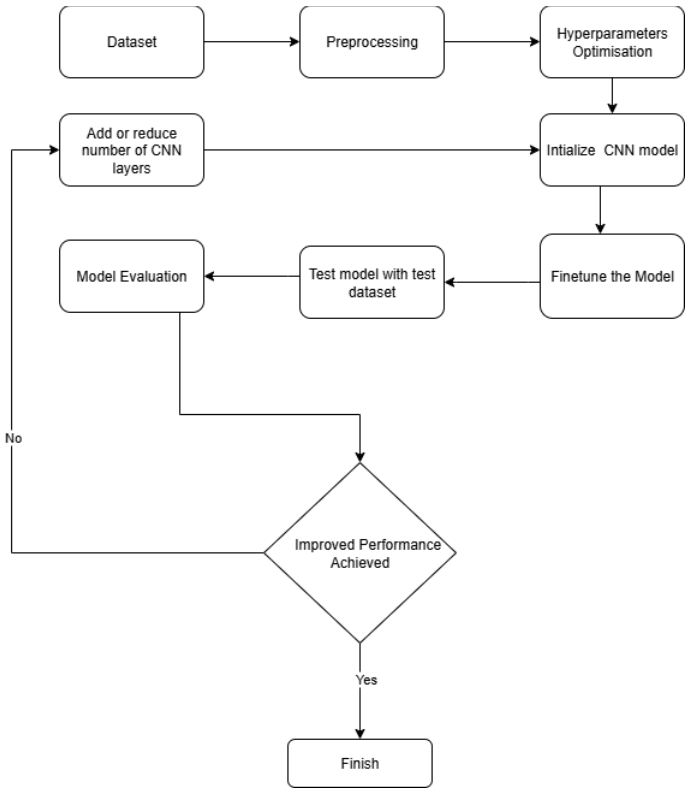


FIGURE 3.1: Methodology

Figure 3.1 illustrates the methodological approach used during CNN model experimentation. The workflow begins with preparing and preprocessing the dataset to ensure high-quality model input. Subsequently, hyperparameter optimization is carried out to fine-tune the model's configuration. The CNN model is initialized with optimal hyperparameters When the outcomes fail to match those documented in existing research, additional refinement takes place, potentially including architectural modifications such as increasing or decreasing the number of CNN layers. Once the model produces results that align with published findings, testing is performed using a separate testing dataset, followed by an evaluation of its performance.

3.1 DataSet

The development of a machine learning model requires the collection of sufficient training and testing [27], This research utilized two video-based facial expression datasets: the Amsterdam Dynamic Facial Expression Set (ADFES) and a Chinese Face Dataset with Dynamic Expressions The ADFES dataset is publicly available at https://aice.uva.nl/research-tools/adfes-stimulus and was used under its academic

research license, which requires proper citation and restricts usage to non-commercial purposes. The Chinese Face Dataset with Dynamic Expressions is publicly available at https://osf.io/7a5fs/ and requires proper citation but has no restrictions on commercial use.

ADFES contains nine basic facial emotions and a neutral expression, whereas the Chinese dataset contains six basic emotions and a neutral expression. This study uses seven basic emotions, as these represent universally recognized emotional states across cultures [16].

3.2 Data Preparation

To prepare the facial expression video datasets, sequences of images were extracted from the videos. A Python script was implemented to process the videos systematically through the main folder. The script extracted the sequence of frames while maintaining the temporal dynamic progression of facial expression.

The naming scheme for the extracted images was based on the initial portion of the original video filename, indicating the associated emotion. For example, if a video was titled was titled `Disgust\_001.mp4`, the frames extracted would be named `Disgust\_001\_frame001.jpg`, `Disgust\_001\_frame002.jpg`. This approach ensured that each image inherently labeled with its respective emotion category. The resulting image sequences were then organized into folders specific to each emotion.

The table below summarizes the number of images per category and provides a sample of extracted images from the videos.

TABLE 3.1: Number of images per emotion category

Emotion	Number of images
Happiness	480
Disgust	855
Fear	856
Neutral	391
Angry	855
Sadness	860
Surprise	840

(A) ADFES images sample

(B) Chinese images sample

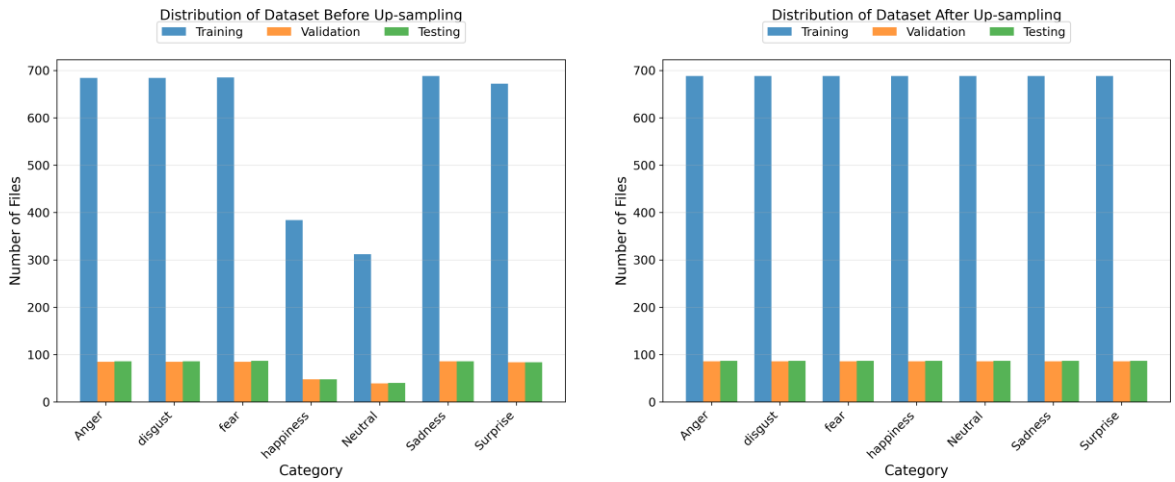
FIGURE 3.2: Sample images from ADFES and Chinese datasets.

3.3 Preprocessing

The pre-processed dataset was systematically divided into three subsets to facilitate proper model training, validation, and evaluation. A custom dataSplitter Python class was implemented to perform the data splitting. The dataset was partitioned using an 80:10:10 split, where 80% was designated for training purposes, 10% for validation, and the remaining 10% for testing. Random shuffling was applied within each class to ensure unbiased sample selection.

The training set was used to optimize the model parameters and learn the underlying patterns in the data. The validation set served as an independent dataset to monitor model performance during training, enabling early stopping to prevent overfitting and assisting in hyperparameter tuning. The test set was reserved exclusively for final model evaluation and was not accessed during the training or validation phases to ensure an unbiased assessment of performance.

After data splitting, random up-sampling of the minority classes was performed to address class imbalance and improve model generalization across all emotion categories.



(A) Distribution of dataset before up-sampling

(B) Distribution of dataset after up-sampling

FIGURE 3.3: Distribution of dataset per category.

Initially, the dataset demonstrated class imbalance, as illustrated in Figure 3.3a. Notably, the happiness and neutral categories have been underrepresented compared to other emotions. To address this issue, random up sampling techniques were employed to up

sample the minority classes. The resulting balanced distribution is shown in Figure 3.3b. The pseudocode in Algorithm 1 outlines the process used to up-sample the minority classes.

Algorithm 1 Random Up-sampling of Minority Classes

Initialize the original dataset D with classes $C_1, C_2, \ldots, C_n$

Determine the target number of samples T for each class (The size of the majority class)

for each class C_i in D **do**

    if $|C_i| < T$ **then**

        Calculate the number of samples to add: $N = T - |C_i|$

        Randomly select N samples from C_i with replacement

        Add the selected samples to C_i

    end if

end for

Combine all classes to form the up-sampled dataset D'

3.4 Hyper-Parameter Tuning using Bayesian Optimisation

In this study, Bayesian optimization was employed for hyperparameter tuning to identify the optimal set of hyperparameters. According to Snoek et al., Bayesian optimization provides an efficient approach to exploring complex hyperparameter spaces and outperforms traditional methods such as grid and random search [60]. The objective in this study was to maximize validation accuracy. By iteratively selecting hyperparameters

expected to improve accuracy, Bayesian optimization efficiently guides the search process, reducing computational resources and accelerating convergence to optimal configurations [55].

The hyperparameter optimization process systematically evaluated three lightweight convolutional neural network architectures pre-trained on ImageNet: MobileNetV2, MobileNetV3 Small, and EfficientNetB0. These architectures were selected for their demonstrated balance between classification accuracy and computational efficiency, making them well-suited for real-time facial expression recognition applications [24, 64]. MobileNetV2 utilizes depthwise separable convolutions, MobileNetV3 Small incorporates neural architecture search optimized for hardware, and EfficientNetB0 employs compound scaling for balanced network dimensions [24, 64].

To leverage pre-trained feature representations, the convolutional layers of each base model were initially frozen. Bayesian optimization was then used to determine the most effective number of layers to unfreeze for fine-tuning. The hyperparameter search space included both architectural and training parameters. Architectural hyperparameters consisted of the number of dense layers added above the CNN feature extractor, the dimensionality of each dense layer, and dropout rates to mitigate overfitting. Training hyperparameters included learning rate and optimizer selection. The loss function was also included in the search to identify the most suitable choice for the task.

The Bayesian optimization procedure ran for approximately 26 hours and 18 minutes on a high-performance computing cluster managed by Simple Linux Utility for Resource Management (SLURM), evaluating 150 trials and successfully identifying an optimal model configuration. MobileNetV2 was selected as the optimal base architecture, achieving superior performance compared to MobileNetV3 Small and EfficientNetB0. This result demonstrates that depthwise separable convolutions with inverted residuals are particularly effective for facial expression recognition, outperforming both the hardware-optimized MobileNetV3 Small and the compound-scaled EfficientNetB0 architectures. Table 3.2 shows the identified optimal hyperparameters.

TABLE 3.2: Hyperparameter Search Results

Parameter	Optimal Value
Best Validation Accuracy	63.46%
Total Elapsed Time	1d 02h 18m 45s
Base Model	MobileNetV2
Layers to Unfreeze	3
Number of Dense Layers	5
Dense Layer 1 Units	448
Dropout 1 Rate	0.150
Dense Layer 2 Units	384
Dropout 2 Rate	0.300
Dense Layer 3 Units	192
Dropout 3 Rate	0.200
Dense Layer 4 Units	512
Dropout 4 Rate	0.300
Dense Layer 5 Units	448
Dropout 5 Rate	0.650
Learning Rate	0.000617
Optimizer	RMSprop
RMSprop Rho	0.850

Loss Function	Focal Loss
Focal Loss Gamma	4.745

After identifying the optimal CNN hyperparameters, the study focused on optimizing the recurrent neural network hyperparameters responsible for capturing temporal dynamics. The previously determined hyperparameters for MobileNetV2 were frozen to preserve the integrity of the learned spatial feature representations, allowing independent tuning of the LSTM.

The LSTM hyperparameters included architectural depth, the number of units per layer, and dropout rates to address potential overfitting. The optimization objective remained consistent with the CNN phase, focusing on maximizing validation accuracy. The tuning process ran for approximately 16 hours and 35 minutes on the same high-performance computing cluster and involved evaluating 150 trials, successfully identifying an optimal LSTM model configuration.

TABLE 3.3: LSTM Hyperparameters

Hyperparameters Number of LSTM layers	Value
LSTM units (layer 1)	320
LSTM dropout (layer 1)	0.20
Best validation accuracy	96.43%

3.5 Proposed Model

Our study proposed two models to classify facial expressions in dynamic images. A standalone MobileNetV2 and a Hybrid MobileNetV2-LSTM are trained separately, and their results are evaluated.

3.5.1 MobileNetV2 Model Architecture

This subsection presents the MobileNetV2 architecture employed in this study. MobileNetV2 was selected for its computational efficiency and its ability to achieve high accuracy with significantly fewer parameters compared to traditional convolutional neural networks. Figure 3.4 illustrates the complete MobileNetV2 architecture, showing the sequence of convolutional layers, inverted residual blocks, and pooling operations that transform the input image into feature representations.

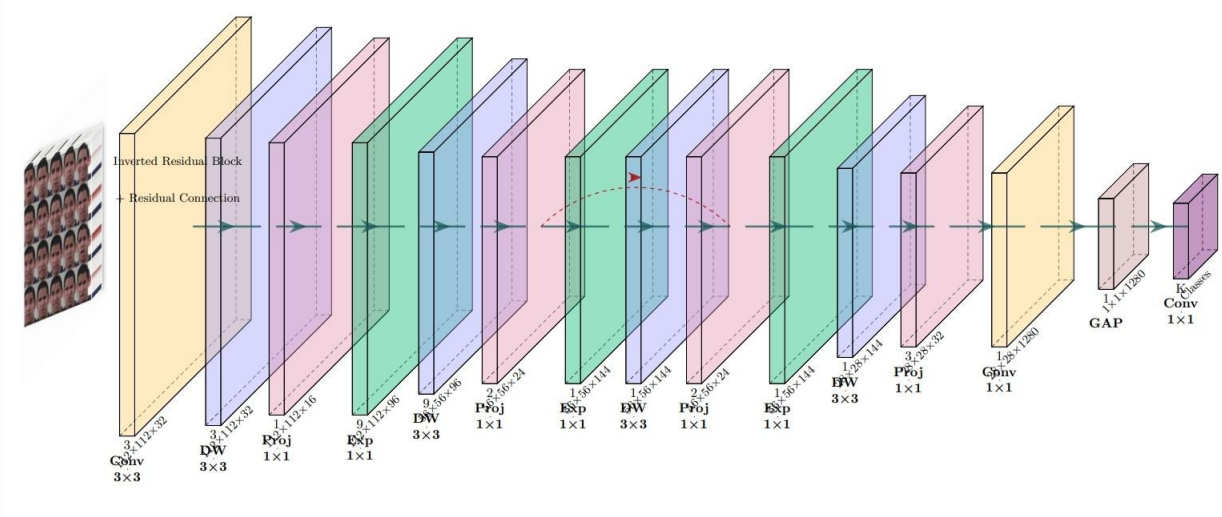


FIGURE 3.4: MobileNetV2 Model Architecture

According to Sinha et al. [59], MobileNetV2 requires input images of fixed dimensions with three RGB channels. Figure 3.4 illustrates how the model employs inverted residual blocks to extract features from input images. It utilizes depthwise separable convolutions, which factorize standard convolution operations into two separate stages: depthwise convolution and pointwise convolution [24]. This architectural design significantly decreases computational complexity while preserving accuracy comparable to conventional convolutional models.

Standard Convolution Mathematically, a standard convolution with kernel **K** of size $D_k \times D_k$, input feature map $\mathbf{F} \in \mathrm{R}^{H\times W\times M}$, and output feature map $\mathbf{G} \in \mathrm{R}^{H\times W\times N}$ can be expressed as shown in Equation (3.1)

$$G_m(i,j) = \sum_{m=1}^{M} \sum_{u=1}^{D_k} \sum_{v=1}^{D_k} K_{k,m}(u,v).F_m(i+u,j+v), \quad (3.1)$$

where M and N denote the number of input and output channels, respectively.

Depthwise Convolution

In depthwise convolution, an individual filter is applied separately to each input channel as shown in Equation (3.2):

$$\hat{F}_m(i,j) = \sum_{u=1}^{D_k} \sum_{v=1}^{D_k} K_m(u,v) \,.\, F_m(i+u, j+v)\,, \tag{3.2}$$

capturing spatial features within each channel without inter-channel mixing.

Pointwise Convolution

The subsequent pointwise convolution (a 1×1 convolution) linearly combines the depthwise outputs across channels as shown in Equation (3.3):

$$G_n(i,j) = \sum_{m=1}^{M} P_{n,m} . \hat{F}_m(i,j), \tag{3.3}$$

where $P_{n,m}$ represents the 1×1 convolution kernel. This operation fuses the spatially filtered outputs from all channels, forming richer feature representations.

Stride Selection and Downsampling

In MobileNetV2, stride values are strategically assigned to specific convolutional layers to balance computational efficiency and feature resolution [59]. Typically, a stride of 2 is applied at the first layer of selected inverted residual blocks, reducing the spatial dimensions of feature maps while increasing the number of channels. The choice of these blocks is guided by the input image size (e.g., 224×224) and the desired feature map resolution for later processing. Layers between these downsampling stages use a stride

of 1 to preserve spatial information, enabling fine-grained feature extraction. This alternating pattern of stride 1 and 2 ensures an optimal trade-off between model compactness, computational efficiency, and representational richness.

Global Average Pooling (GAP)

After the final convolutional block, a Global Average Pooling (GAP) layer reduces each feature map to a single representative value per channel as shown in Equation (3.4):

$$Z_{c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j),} \tag{3.4}$$

where $x_c(i,j)$ denotes the activation value at position (i,j) in the c-th feature map, and H and W are the spatial dimensions. GAP summarizes spatial information and produces a compact feature vector, helping minimize overfitting.

Dropout Regularization

To regularize the model, a Dropout layer randomly deactivates neurons during training with probability p as shown in Equation (3.5):

$$\mathrm{y_{i=\frac{x_i m_i}{1-p}}},m \sim \mathrm{Bernoulli}(1-p), \tag{3.5}$$

Where x_i is the input activation and m_i is a binary mask sampled from a Bernoulli distribution. The scaling factor $\frac{1}{1-p}$ maintains consistent activation magnitude during inference.

Fully Connected Layer and Softmax Classification

Finally, a fully connected (Dense) layer maps the pooled features to emotion categories through a linear transformation followed by a softmax activation as shown in Equation (3.6,37):

$$\hat{\mathrm{Y}} = \mathrm{softmax}(Wz + \mathrm{b}), \tag{3.6}$$

$$\text{softmax}(y_i) = \frac{e^{y_i}}{\sum_{j=1}^{k} e^{y_j}}, \tag{3.7}$$

where **W** and **b** are the trainable weight matrix and bias vector, **z** is the GAP output vector, and K is the total number of emotion classes. The softmax function converts logits into normalized probabilities for classification,

3.5.2 MobileNetV2-LSTM Model Architecture

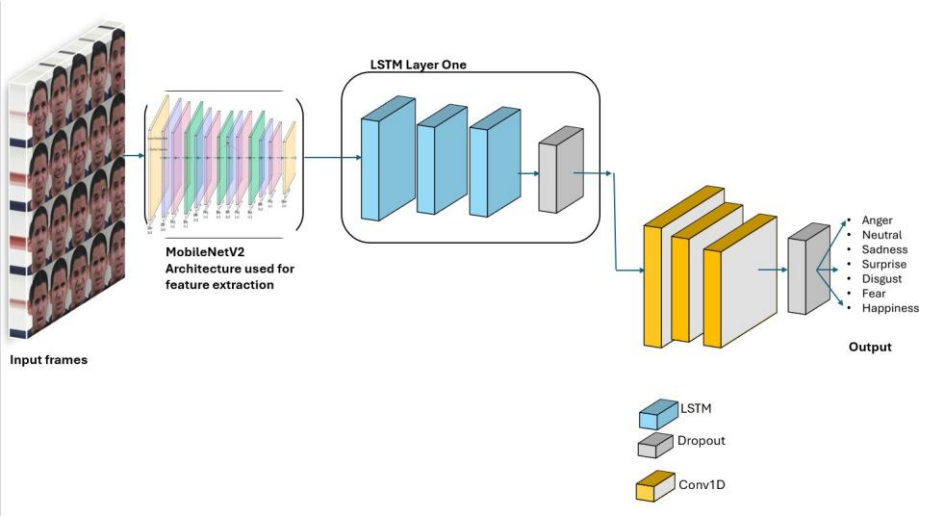


FIGURE 3.5: MobileNetV2-LSTM Model Architecture

Figure 3.5 above shows a fine-tuned version of MobileNetV2 [24, 59] with a one layer LSTM to capture temporal features. Our architecture uses the previously discussed MobileNetV2 for efficient feature extraction. After the feature extraction from MobileNetV2's inverted residual blocks, the extracted features are fed into LSTM layers. The LSTM layer models dynamic changes in the sequence of images, capturing the temporal relationship between the sequential frames in the data [8].

Mathematically, the LSTM unit at time step t is defined by the following equations (38-3.13):

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f), \tag{3.8}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), \quad (3.9)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c), \quad (3.10)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \quad (3.11)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o), \quad (3.12)$$

$$h_t = o_t \odot \tanh(c_t), \quad (3.13)$$

where x_t is the input at time t (the features from MobileNetV2), h_t is the hidden state, c_t is the cell state, f_t, i_t, o_t are the forget, input, and output gates, respectively,

and $\odot$ denotes element-wise multiplication. W_*, U_*, and b_* are trainable weight matrices and biases.

The forget gate f_t determines which information from the previous cell state c_{t-1} is retained. The input gate i_t controls how much of the new candidate information $\tilde{c}_t$ is added to the cell state. The cell state c_t stores the long-term memory, updated by combining retained old information and new input. The output gate o_t regulates which part of the cell state is output as the hidden state h_t, representing the LSTM's output at the current time step.

The combination of MobileNetV2's efficient spatial feature extraction and LSTM's temporal modeling creates a robust architecture suitable for real-time sequential data processing. After the LSTM layer, a dropout layer is introduced to prevent overfitting and improve generalization. 1D convolutional layers are employed to refine the temporal features extracted by the LSTM, allowing the model to capture local temporal patterns in the sequence. The final output layer with softmax activation classifies different facial expression emotions based on both spatial features of MobileNetV2 and temporal dependencies captured by the LSTM network. This hybrid architecture leverages the strengths of both convolutional networks for spatial feature extraction and recurrent networks for temporal sequence modeling, while

maintaining the computational efficiency characteristic of MobileNetV2.

3.6 Model Evaluation

To evaluate the performance of our facial expression models, our study used both classification accuracy metrics and computational efficiency measures [41, 35]. This approach ensures the model demonstrates not only high predictive performance but also practical feasibility for deployment in real-time applications where latency limitations are a key consideration [77, 79]. T he predictive performance was evaluated using the following metrics: accuracy, precision, recall, and F1-score. Additionally, a confusion matrix was generated to provide a visual representation of prediction distributions and identify commonly confused emotion classes. Given the real-time requirements of facial expression recognition applications, computational efficiency was evaluated through system throughput (frames per second) and per-frame inference time metrics.

Accuracy: Accuracy denotes the ratio of correctly classified instances to the overall number of instances. In the following formulas, TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. The equation for computing accuracy is presented in Equation (3.14).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \tag{3.14}$$

Precision: Precision quantifies the ratio of correctly identified positive instances among all instances classified as positive. It assesses the model's capability to minimize false positives. Equation (3.15) illustrates how precision is calculated.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3.15}$$

Recall: Recall assesses the model's ability to correctly identify True Positives. The method for calculating recall is provided in Equation (3.16)

$$\text{Recall } \frac{TP}{TP+FN} \tag{3.16}$$

F1-score: The F1-score reflects the harmonic balance between Recall and Precision. It is computed using Equation (3.17).

$$\text{F1} - \text{score } = \frac{2 \times \text{Precision} \times \text{Recall}}{TP + \text{Precision} + \text{Recall}} \tag{3.17}$$

3.7 Computational Efficiency Evaluation

According to Zhang et al., the constraints inherent in real-time facial recognition applications necessitate the simultaneous achievement of computational efficiency and high predictive accuracy [77, 79]. Our study measured system throughput and inference time. System throughput was measured for both single images and sequence-based images, depending on the input type. For single images, it was defined as the number of individual images processed per second, while for sequence-based it was the number of video frames processed per second during inference. Additionally, per-sample inference time was calculated as the average temporal cost required to process individual inputs. The mathematical definitions for these efficiency metrics are presented below. Where N_{images} and N_{frames} represent the total number of images and video frames processed, respectively, and T_{total} denotes the total inference time in seconds.

For single-image processing mathematically is defined by the following equations (3.18-3.19):

$$\text{Throughput (IPS)} = \frac{T_{total}}{N_{images}} \tag{3.18}$$

$$\text{Inference time per image} = \frac{T_{total}}{N_{images}}, \quad (3.19)$$

For frame processing mathematically is defined by the following equations (3.20-3.21):

$$\text{Throughput}(FPS) = \frac{N_{frames}}{T_{total}}, \quad (3.20)$$

$$\text{Inference time per frame} = \frac{N_{total}}{T_{frames}}, \quad (3.21)$$

3.8 Tools

The study employs the Python programming language for model development, frame extraction and face detection, leveraging its extensive machine learning ecosystem and ease of use for rapid prototyping. It also uses OpenCV for image and video preprocessing. TensorFlow and Keras are used to design, train, and evaluate deep learning models, such as MobileNetV2, and MobileNetV2 for facial expression recognition. GPU acceleration is utilized to enhance computational efficiency and reduce training time

3.9 Ethical Considerations

This research utilizes publicly accessible datasets; therefore, no direct engagement with human subjects or gathering of personal information was required. At the University of Venda, all research proposals must undergo ethical review and receive approval prior to the commencement of the study. Accordingly, the proposal for this study was initially submitted to the Department of Mathematical and Computational Sciences, where it was accepted with recommendations for revisions. After incorporating the suggested revisions, the proposal was submitted to the Faculty Research and Innovation Committee (FRIC) via the Department's research ethics representative. This ethical clearance process ensures that the research adheres to the university's ethical guidelines and maintains research integrity. Ethical clearance was subsequently granted by FRIC, and the certificate is provided Appendix A. Additionally, the study underwent a plagiarism check using Turnitin software. These steps collectively uphold the integrity and ethical standards of research conducted at the University of Venda.

Although this study will not be deployed in a real surveillance environment, facial expression recognition technology raises considerable ethical concerns, including privacy risks, potential for misuse, and algorithmic bias. Such systems may be applied in real-world contexts without individuals' consent, raising concerns regarding data security and fairness. To ensure responsible research practice, this study focused on model development and evaluation using publicly available datasets and adhered to ethical AI guidelines to promote fairness, accountability, and transparency.

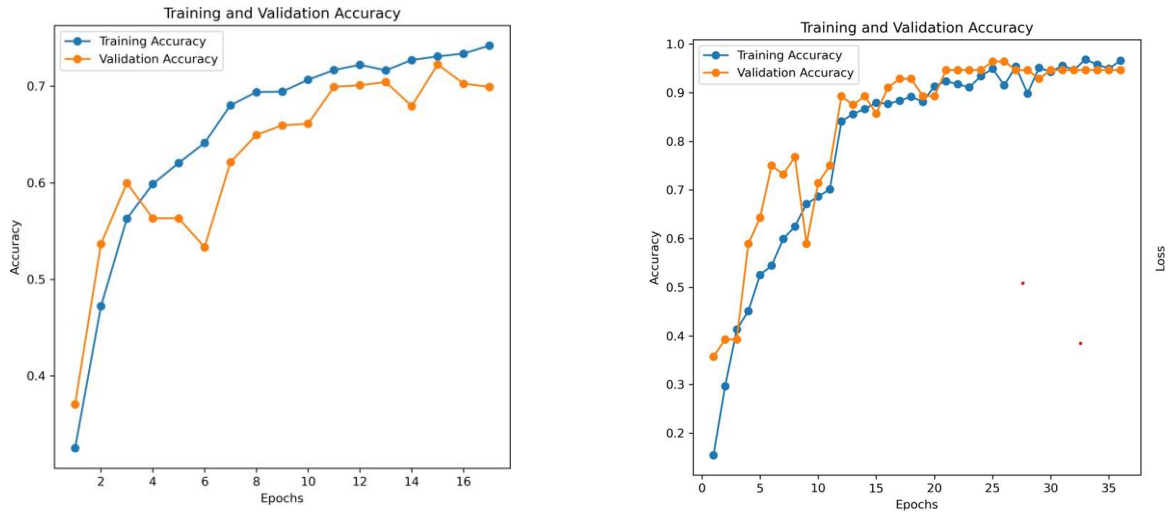
3.10 Conclusion

In conclusion, the methodology of the study presents a comprehensive framework for developing and evaluating facial expression recognition models. Python programming language, along with its machine learning libraries OpenCV, TensorFlow, and Keras, was used to ensure efficient preprocessing, model training, and evaluation. Deep learning architectures, including MobileNetV2 and MobileNetV2–LSTM, combined with GPU acceleration, enabled accurate and computationally efficient recognition of facial expressions. This approach ensures that the research objectives are effectively addressed while maintaining scientific rigor and adherence to ethical standard.

Chapter 4 Results

This chapter presents the results obtained from the development and evaluation of facial expression recognition models as outlined in the methodology. The aim is to assess the performance of the implemented models, namely MobileNetV2 and the MobileNetV2–LSTM hybrid, on the selected datasets. The results include quantitative evaluations such as accuracy, precision, recall, and F1-score, as well as qualitative observations from model predictions on sample images and video frames.

4.1 Model Training

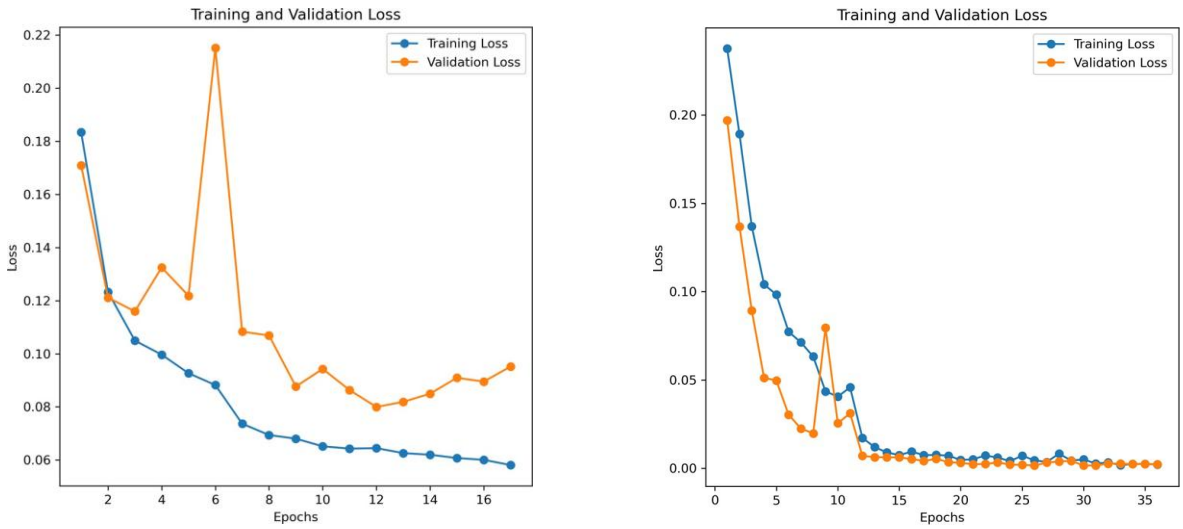


(A) MobileNetV2 Training Accuracy

(B) MobileNetV26LSTM Training Accuracy

FIGURE 4.1: Accuracy Graphs

Figure 4.1 above illustrates the training and validation accuracy curves for the MobileNetV2 model (Figure 4.1a) and the MobileNetV2-LSTM model (Figure 4.1b) over their respective training epochs. The graphs indicate differences in convergence behaviour and generalization performance between the two models, which are discussed in detail in Chapter 5.



(A) MobileNetV2 Training Loss

(B) MobileNetV2LSTM Training Loss

FIGURE 4.2: Loss Graphs

Figure 4.2 above illustrates the training and validation loss curves for the MobileNetV2 model (Figure 4.2a) and the MobileNetV2-LSTM model (Figure 4.2b) over their respective training epochs. The graphs indicate differences in loss reduction and stability between the two models, which are discussed in detail in Chapter 5.

4.2 Evaluation

This section presents the experimental findings comparing two deep learning architectures, MobileNetV2 and MobileNetV2 integrated with LSTM, for emotion recognition within a surveillance system. Each model's performance is assessed through confusion matrices and classification reports.

TABLE 4.1: MobileNetV2 Classification Report

Class	Precision	Recall	F1-Score	Support
Anger	0.81	0.48	0.60	87
Neutral	0.88	0.93	0.91	87
Sadness	0.83	0.63	0.72	87

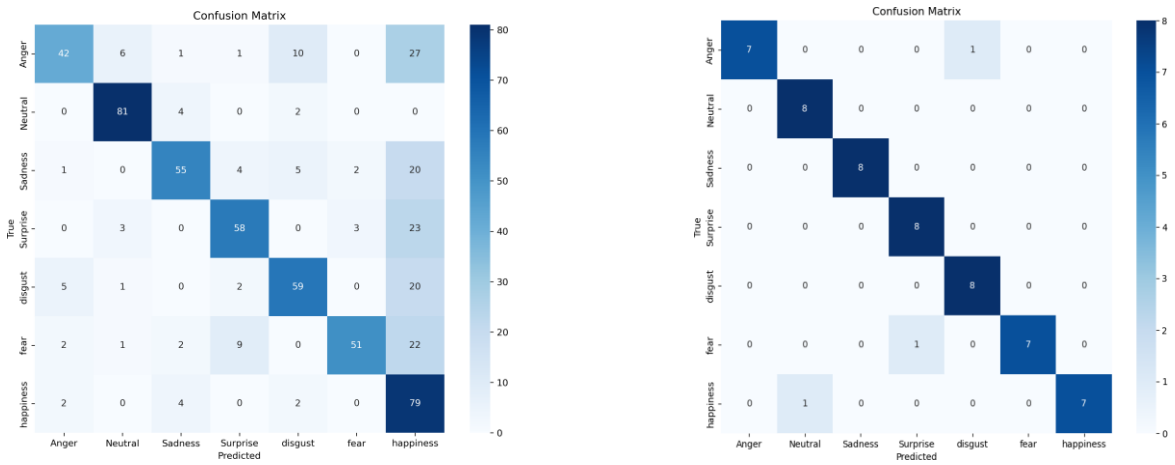
Surprise	0.78	0.67	0.72	87
Disgust	0.76	0.68	0.72	87
Fear	0.91	0.59	0.71	87
Happiness	0.41	0.91	0.57	87
Accuracy			0.70	609
Macro Avg	0.77	0.70	0.71	609
Weighted Avg	0.77	0.70	0.71	609

Table 4.1 above presents the classification report for the MobileNetV2 model across seven facial expression classes. The results show variation in precision, recall, and F1-score across classes, with Neutral achieving the highest F1-score of 0.91 and Happiness the lowest at 0.57, resulting in an overall accuracy of 70%. A detailed analysis of these results is provided in Chapter 5.

TABLE 4.2: MobileNetV2-LSTM Classification Report

Class	Precision	Recall	F1-Score	Support
Anger	1.00	0.88	0.93	8
Neutral	0.89	1.00	0.94	8
Sadness	1.00	1.00	1.00	8
Surprise	0.89	1.00	0.94	8
Disgust	0.89	1.00	0.94	8
Fear	1.00	0.88	0.93	8
Happiness	1.00	0.88	0.93	8
Accuracy			0.95	56
Macro Avg	0.95	0.95	0.95	56
Weighted Avg	0.95	0.95	0.95	56

Table 4.2 above presents the classification report for the MobileNetV2-LSTM model across seven facial expression classes. The results demonstrate consistently high precision, recall, and F1-scores across all classes, with Sadness achieving a perfect F1-score of 1.00 and an overall accuracy of 95%. A detailed analysis of these results is provided in Chapter 5.



(A) MobileNetV2 Confusion Matrix

(B) MobileNetV2LSTM Confusion Matrix

FIGURE 4.3: Confusion Matrix

Figure 4.3 above presents the confusion matrices for the MobileNetV2 model (Figure 4.3a) and the MobileNetV2-LSTM model (Figure 4.3b), illustrating the classification performance across all seven facial expression classes. The matrices highlight differences in misclassification patterns between the two models, which are discussed in detail in Chapter 5.

4.3 Real-World Challenges Evaluation

TABLE 4.3: MobileNetV2-LSTM Classification Report (Occluded)

Class	Precision	Recall	F1-Score	Support
Anger	0.30	0.38	0.33	8
Neutral	1.00	0.12	0.22	8
Sadness	0.50	0.12	0.20	8
Surprise	1.00	1.00	1.00	8
Disgust	0.32	1.00	0.48	8
Fear	1.00	0.75	0.86	8

Happiness	1.00	0.50	0.67	8
Accuracy			0.55	56
Macro Avg	0.73	0.55	0.54	56
Weighted Avg	0.73	0.55	0.54	56

Table 4.3 above presents the classification report for the MobileNetV2-LSTM model tested on occluded images, achieving an overall accuracy of 55%. A detailed analysis is provided in Chapter 5

TABLE 4.4: MobileNetV2-LSTM Classification Report (Blurred)

Class	Precision	Recall	F1-Score	Support
Anger	0.75	0.38	0.50	8
Neutral	1.00	0.38	0.55	8
Sadness	0.57	1.00	0.73	8
Surprise	1.00	0.62	0.77	8
Disgust	0.62	1.00	0.76	8
Fear	0.73	1.00	0.84	8
Happiness	1.00	0.75	0.86	8
Accuracy			0.73	56
Macro Avg	0.81	0.73	0.71	56
Weighted Avg	0.81	0.73	0.71	56

Table 4.4 above presents the classification report under blurred conditions, showing an improved overall accuracy of 73% compared to occlusion. A detailed analysis is provided in Chapter 5.

TABLE 4.5: MobileNetV2-LSTM Classification Report (Occluded + Blurred)

Class	Precision	Recall	F1-Score	Support
Anger	0.25	0.25	0.25	8
Neutral	0.00	0.00	0.00	8
Sadness	0.50	0.25	0.33	8
Surprise	1.00	0.62	0.77	8
Disgust	0.25	1.00	0.40	8
Fear	0.75	0.38	0.50	8
Happiness	1.00	0.38	0.55	8
Accuracy			0.41	56
Macro Avg	0.54	0.41	0.40	56
Weighted Avg	0.54	0.41	0.40	56

Table 4.5 above presents the classification report under combined occlusion and blurring conditions, resulting in the lowest overall accuracy of 41% across all three conditions. A detailed analysis is provided in Chapter 5.

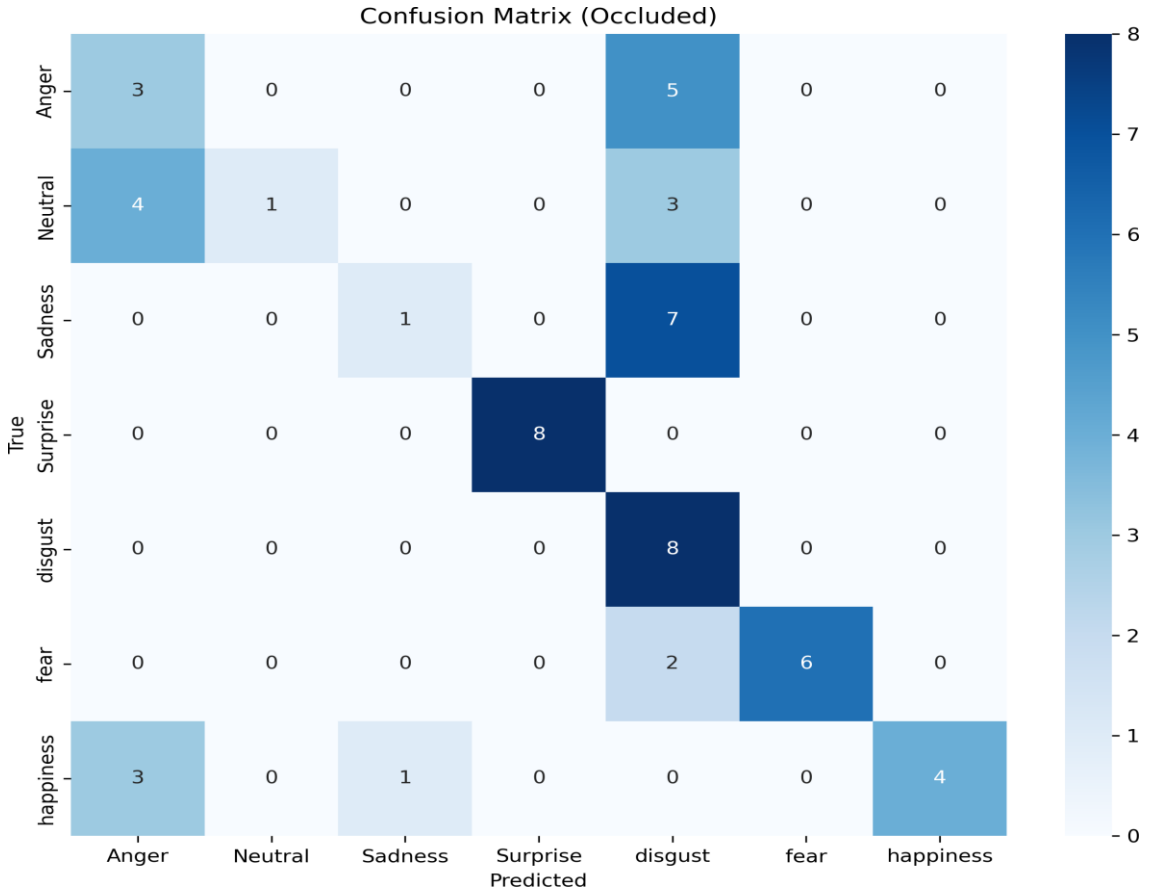


FIGURE 4.4: Confusion Matrix for Occluded Images using MobileNetV2-LSTM

Figure 4.4 above presents the confusion matrix for the MobileNetV2-LSTM model evaluated on occluded images, illustrating the classification performance across all seven facial expression classes. A detailed analysis is provided in Chapter 5.

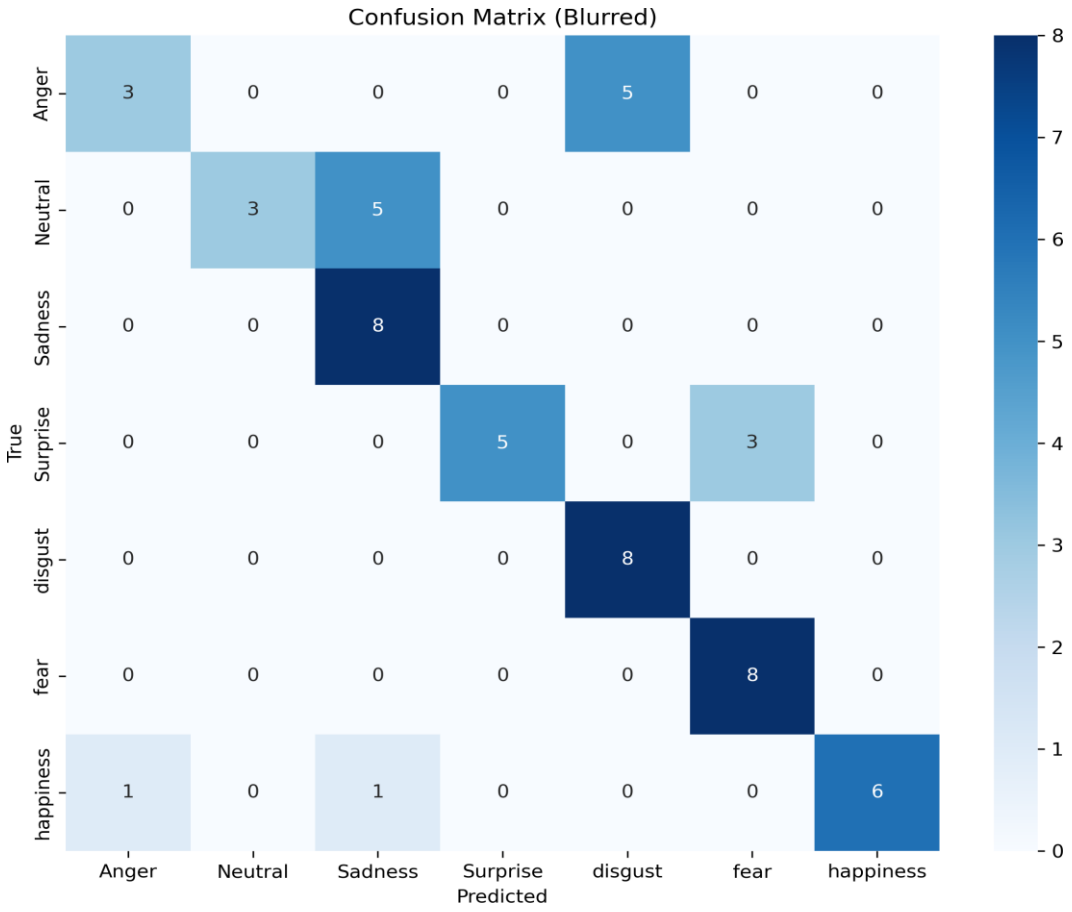


FIGURE 4.5: Confusion Matrix for Blurred Images using MobileNetV2- LSTM

Figure 4.5 above presents the confusion matrix for the MobileNetV2-LSTM model evaluated on blurred images, illustrating the classification performance across all seven facial expression classes. A detailed analysis is provided in Chapter 5.

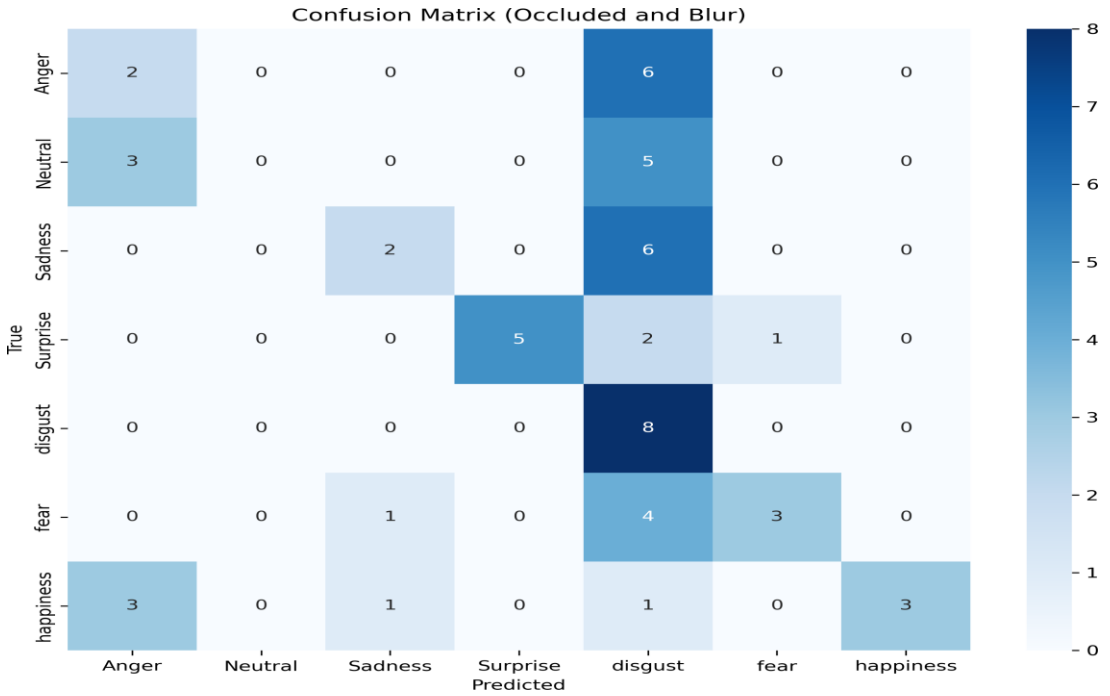


FIGURE 4.6: Confusion Matrix for Occluded + Blurred Images using MobileNetV2-LSTM

Figure 4.6 above presents the confusion matrix for the MobileNetV2-LSTM model evaluated on combined occluded and blurred images, illustrating the classification performance across all seven facial expression classes. A detailed analysis is provided in Chapter 5.

4.4 Computational Evaluation

This subsection presents the computational performance metrics of the proposed models, specifically focusing on inference time and system throughput. The evaluation compares the MobileNetV2 model, which processes individual images independently, against the MobileNetV2–LSTM model, which processes sequences of frames to leverage temporal information.

TABLE 4.6 : Computational Evaluation of MobileNetV2 and MobileNetV2-LSTM Models

Metric	**MobileNetV2 (Image-level)**	**MobileNetV2-LSTM (Sequence-level)**
Average inference time	0.0718 seconds/image	0.2321seconds/sequence
Average inference time per frame	—	0.0232 seconds/frame
System throughput (FPS)	13.93 images/second	4.31 sequences/second
Equivalent FPS (images/sec)	—	43.09 images/second
Sequence length	—	10 frames/sequence

Table 4.6 above presents the computational evaluation of both models, comparing inference time, system throughput, and frames per second. A detailed analysis of these results is provided in Chapter 5.

4.5 Conclusion

The experimental results for MobileNetV2 and MobileNetV2–LSTM demonstrate differences in performance, computational efficiency, and applicability for surveillance-based facial expression recognition. The MobileNetV2–LSTM model achieved higher accuracy, precision, recall, and F1-score compared to MobileNetV2, with confusion matrices showing improved classification across most emotion classes. These results indicate that incorporating temporal modeling with LSTM provides measurable performance gains while maintaining real-time efficiency. However, the model's accuracy declined significantly under occlusion and blur, highlighting limitations for real-world deployment. Ethical considerations, including privacy, consent, and bias mitigation, were also addressed.

The next chapter discusses the results presented in this chapter providing critical analysis and interpretation.

Chapter 5 Discussion

This chapter presents the discussion and interpretation of the models' effectiveness in recognizing facial expressions.

5.1 Models Training

Figure 4.1 shows the accuracy graphs for the two trained models. Figure 4.1a shows the training and validation accuracy patterns for the MobileNetV2 model, which was trained using early stopping to avoid overfitting. The model's accuracy increases sharply during the initial epochs, but by epoch 8, validation accuracy plateaus and then slightly declines. This divergence indicates overfitting, and early stopping was triggered at epoch 17. The final training accuracy reached 74.49%, while validation accuracy declined, highlighting the generalization gap.

Figure 4.1b illustrates the MobileNetV2-LSTM model's performance. The training and validation accuracy progressed more steadily and remained closely aligned. By epoch 36, training accuracy reached 95.85%, and validation accuracy reached 94.64%, showing a minimal generalization gap of 1.21%. This suggests that the LSTM component contributed to regularization, helping the model generalize better to unseen data.

Figure 4.2 shows the training and validation loss for the two models trained. Figure 4.2a shows the loss trends for the MobileNetV2 model. The training loss de- decreases consistently as the model learns throughout, indicating effective learning. However, it is evident that the validation loss starts to vary after several epochs, with a notable spike occurring around epoch 6, where the validation loss sharply increases from 0.12 to 0.21. This spike suggests the model encountered difficulty generalizing at this point in training. Following the spike, the validation loss recovers but continues to fluctuate, suggesting that while the model continues to perform better on the training data, it does not generalize as effectively to the validation data. Furthermore, this confirms the overfitting observed in the loss graphs.

Figure 4.2b illustrates the loss trends for the MobileNetV2-LSTM hybrid model,

demonstrating markedly different characteristics compared to the standalone architecture. Both training and validation loss curves show smooth, consistent decline throughout the extended training period. The absence of validation loss plateaus indicates superior generalization.

5.2 Performance Evaluation

5.2.1 Classification Reports

Table 4.1 presents the performance metrics for MobileNetV2 across different emotions. An overall accuracy of 70 percent demonstrates that the model accurately classified 70 percent of the emotions within our dataset. The model exhibits strong precision for fear, correctly recognizing it 91 percent of the time when predicting fear. Conversely, the low recall of 59 percent indicates that it overlooks numerous fear instances, implying that while the model infrequently misidentifies other emotions as fear, it frequently fails to detect fear when it occurs.

Anger expressions present a significant challenge for the model. Although precision is relatively high at 81 percent, meaning predictions of anger are often correct, recall is extremely low at 48 percent, indicating that the model fails to detect most anger instances. Sadness detection demonstrates good precision at 83 percent, but recall remains low at 63 percent. These findings for negative emotions are consistent with previous studies, which show that negative emotions, particularly fear and anger, are more difficult to detect than positive emotions due to their subtle and variable expression patterns [74, 26, 47]. Surprise shows moderate precision at 78 percent but a higher recall of 67 percent, indicating that the model is reasonably balanced in classifying expressions as surprise, correctly identifying 67 percent of all surprise cases. Disgust exhibits good precision at 76 percent and moderate recall at 68 percent, suggesting that some instances of disgust are missed. Neutral detection demonstrates strong precision at 88 percent and excellent recall at 93 percent, reflecting the model's effectiveness in recognizing neutral expressions. It is important to note that the neutral class originally had a low number of samples, and upsampling techniques were applied to address this imbalance, which likely contributed to the improved performance.

Happiness shows the highest recall among all emotions, correctly identifying 91 percent of happy expressions. However, the precision is critically low at 41 percent, indicating frequent misclassification of other emotions as happiness. This pattern is attributable to

upsampling of the happiness class due to its minority status in the original dataset.

The MobileNetV2-LSTM model achieved an overall accuracy of 95 percent, representing a 25 percent improvement over the baseline MobileNetV2 model. This improvement was notably evident in the detecting sadness and surprise, as reflected F1-scores that improved to 1.00 and 0.94, respectively. The model demonstrated strong performance across all emotion categories, with macro-average precision and recall both reaching 0.95. This represents a substantial improvement over existing dynamic emotion recognition systems, which typically report macro-average precision and recall in the range of 0.38 to 0.59 [74, 47].

Notably, sadness is recognized perfectly, with both precision and recall at 1.00, indicating flawless classification. Anger, fear, and happiness achieve perfect precision at 1.00 and high recall at 88 percent, showing robust performance with minimal misclassifications. Neutral expressions maintain strong performance, with 89 percent precision and 100 percent recall, reflecting the LSTM layers' ability to distinguish temporal patterns and reduce confusion from upsampling. Surprise and disgust both achieve 89 percent precision with 100 percent recall. Overall, the sequence processing capability of the LSTM enhances the model's ability to differentiate emotions effectively, with all emotions achieving F1-scores of 93 percent and higher.

5.2.2 Confusion Matrix

Figure 4.3 presents the confusion matrices for the proposed models, providing detailed insights into emotion recognition performance across classes.

For the MobileNetV2 model (Figure 4.3a), happiness exhibited optimal recognition performance with 79 accurate classifications, illustrating model's robust capacity in identifying positive expressions. However, happiness is also the most frequent source of misclassification across other emotions, reflecting bias introduced by upsampling of minority classes. Anger proves challenging, with 42 correct predictions, 27 misclassified as happiness, and 10 as disgust, likely due to overlapping facial muscle activations. Surprise achieves 58 correct predictions, but 23 instances are misclassified as happiness, indicating similar bias. Fear shows 51 correct predictions, with 22 misclassified as happiness, while sadness and disgust also experience confusion with happiness. Neutral

expressions achieve 81 correct predictions, indicating that upsampling was effectively managed for this class.

In contrast, the MobileNetV2-LSTM model (Figure 4.3b) demonstrates notable improvements. The LSTM layers enhance the model's ability to differentiate emotions by capturing temporal dynamics. Happiness achieves near-perfect recognition with 7 out of 8 correct predictions. Surprise, sadness, and disgust achieve perfect classification, and neutral expressions are also perfectly recognized. Fear and anger exhibit minimal misclassification. These results highlight that incorporating temporal modelling improves classification consistency and mitigates biases introduced by up sampling.

5.3 Real-World Challenges Evaluation

To assess the practicality of our model for deployment in real-world surveillance environments, its performance was evaluated under common visual degradations, including blur and occlusion [45, 46]. The classification reports in Tables 4.3–4.5 and the corresponding confusion matrices in Figures 4.4–4.6 summarize the model's performance under these conditions.

For occluded images, overall accuracy dropped to 55%, with neutral and sadness classes exhibiting notable misclassifications. Precision and recall remained high for fear and surprise, suggesting that the model can still identify expressions with distinctive features even under partial obstruction. This indicates that MobileNetV2-LSTM effectively captures high-saliency facial cues, though it struggles with subtler expressions that rely on the full facial context.

Under blurred conditions, accuracy improved to 73%, with better recall across most classes compared to the occlusion scenario. This improvement highlights the benefit of the LSTM's temporal modeling, as the sequential processing of frames allows the model to extract consistent motion patterns despite spatial degradation.

When both occlusion and blur were applied simultaneously, overall accuracy declined further to 41%, with macro and weighted averages reflecting the compounded challenges of these distortions. Misclassifications were particularly severe for neutral and anger classes, demonstrating the model's sensitivity to multiple visual artifacts. Nevertheless,

emotions such as surprise and happiness were relatively well recognized, emphasizing the model's reliance on prominent temporal and spatial cues.

Overall, these evaluations indicate that while MobileNetV2-LSTM performs effectively under moderate real-world distortions, its robustness diminishes under extreme conditions. Compared to purely convolutional models [45, 46], hybrid CNN-LSTM architectures are generally more resilient to dynamic challenges. However, further improvements could be achieved through strategies such as data augmentation, multi-scale feature extraction, and attention-based mechanisms [59]. The findings suggest that the proposed model is suitable for deployment in environments with moderate visual degradation, such as low-light surveillance or partial occlusion, but additional enhancements would be needed for highly challenging scenarios.

5.4 Computational Evaluation

Table 4.6 summarizes the average inference times per image and per sequence, along with the corresponding frames-per-second (FPS) throughput rates. These metrics provide important insight into the suitability of the models for real-time facial expression recognition in video streams.

The objective of this study was to develop an intelligent surveillance system using a deep neural network, specifically a CNN and a hybrid CNN-RNN, trained on dynamic images to accurately capture facial expression changes in real-time surveillance scenarios. The results demonstrate that this objective was successfully achieved. The proposed hybrid model achieved 95% accuracy for facial expression recognition while maintaining real-time processing at 43.09 FPS during sequence-based processing. The integration of the LSTM component enabled effective capture of temporal dynamics in facial expressions, addressing limitations inherent in static image-based approaches.

According to Matsumoto and Hwang, negative emotional micro expressions are reliable indicators of concealed malicious intentions [81]. Their study demonstrates that Towards Facial Expression Analysis for Enhanced Threat Detection 57 individuals with malicious intent exhibit significantly more negative expressions, such as anger, disgust, fear, and sadness, within 0.40–0.50 s. Our temporal dynamic approach captures these negative

emotional patterns for threat detection in surveillance systems.

TABLE 5.1: Comparison of accuracy across different methods.

Authors	Model Used	Accuracy (%)
Current Study	CNN-RNN	95%
Yin et al.[74]	CNN-RNN	44%
Rangulov and Fahim [47]	CNN-RNN	50%
Fan et al. [20]	CNN-RNN-C3D	46%
Lu et al [37]	CNN-RNN	49%
Lu et al[37]	CNN-RNN-C3D	60%

Table 5.1 presents a comparative evaluation of the proposed CNN-RNN model against previously reported approaches in the literature. The model developed in this study achieved an overall accuracy of 95%, outperforming existing CNN-RNN-based architectures [74, 20, 47, 37]. Previous works, including Yin et al. [74], Rangulov and Fahim [47], and Lu et al. [37], reported lower classification accuracies of 44%, 50%, and 49%, respectively, highlighting the limited generalization capacity of earlier CNN-RNN models, particularly in modeling dynamic expression sequences. CNN-RNN-C3D architectures showed modest improvement, 46% by Fan et al. [20] and 60% by Lu et al. [37] but still performed considerably below the model proposed in this study. Furthermore, it is important to note that although similar CNN-RNN architectures were employed, differences in dataset characteristics, preprocessing techniques, and hyperparameter tuning across studies explain the observed performance variations.

The superior performance of the proposed model is attributed to several architectural and training refinements. Bayesian optimization was used to systematically tune hyperparameters for both the CNN and recurrent components, enabling faster training and identification of optimal configurations. The use of temporal sequences as input

allowed the model to capture dynamic changes across frames effectively. Additionally, regularization techniques, including dropout and early stopping, contributed to improved generalization and robustness.

While CNN-RNN-C3D architecture has demonstrated potential in capturing spatio-temporal features, the findings here indicate that carefully optimized CNN-RNN combinations can achieve high performance at a reduced computational cost. These results underscore the relevance of lightweight, sequence-aware architectures for real-time emotion recognition tasks in practical surveillance applications.

5.5 Ethical Considerations and Guidelines for Responsible Use

While this research uses publicly available datasets, the application of facial expression recognition in surveillance systems raises critical cybersecurity concerns [63]. Emotion recognition data represents highly sensitive personal information that, once compromised, cannot be changed like traditional credentials. The technical capabilities demonstrated in this study could be applied in real-world surveillance systems that process biometric information without the consent of the data subject, raising important questions of privacy, fairness, and accountability [63].

The observed performance decline under real-world conditions, particularly the drop in accuracy from 95% to 41% under combined occlusion and blur scenarios, highlights potential vulnerabilities that could be exploited to manipulate emotion classification results. These findings emphasize the need for robust cybersecurity measures, transparent governance, and privacy-preserving system architectures in future surveillance deployments. Such measures are essential to ensure compliance with regulatory frameworks, including the South African Protection of Personal Information Act (POPIA) [13].

To address these concerns, we propose the following guidelines for the responsible deployment of facial expression recognition systems, drawing on established ethical frameworks for biometric technologies [7] and privacy-preserving machine learning practices [72]. Data governance should include obtaining explicit informed consent from

individuals prior to data collection [28], implementing secure storage with encryption and access controls, establishing clear data retention and deletion policies, and ensuring full compliance with POPIA and relevant international regulations. Technical safeguards should include regular bias audits across demographic groups [67], setting minimum performance thresholds before deployment, and implementing privacy-preserving techniques such as data anonymization and differential privacy [23]

5.6 Conclusion

This chapter evaluated the MobileNetV2 and MobileNetV2-LSTM models, examining their performance, robustness, computational efficiency, and ethical considerations in surveillance-based facial expression recognition. The MobileNetV2-LSTM model outperformed the baseline in controlled settings, effectively capturing temporal dynamics while maintaining real-time efficiency. However, its accuracy declined significantly under occlusion and blur, highlighting limitations for real-world deployment. Ethical guidelines for responsible use were also discussed. The next chapter presents the study's overall conclusions, contributions, and recommendations for future research

Chapter 6 Conclusion and Future Research Work

This chapter serves as the culmination of our study on deep learning approaches for facial expression recognition in surveillance systems. We utilized the Amsterdam Dynamic Facial Expression Set (ADFES) and the Chinese Face Dataset with Dynamic Expressions to enhance the real-world applicability of our findings, providing a basis for evaluating both spatial and temporal modeling of facial expressions.

Our research examined CNN and CNN-LSTM architectures, highlighting their ability to capture spatial features through MobileNetV2 and temporal dependencies through LSTM layers. The dynamic characteristics of ADFES, which include face-forward and head-turning variations, combined with the Chinese dataset, allowed the models to learn robust representations from diverse expressions. The evaluation results demonstrated strong classification performance and computational efficiency, achieving high accuracy while maintaining real-time feasibility using dynamic sequences rather than static images.

In this concluding chapter, we reflect on the main contributions of our study, discuss practical implications for surveillance applications, and identify challenges related to robustness, demographic diversity, and ethical deployment. Finally, we outline future directions aimed at further enhancing model accuracy, generalizability, and real-time deployment capabilities, including the incorporation of more diverse ethnicity datasets and privacy-preserving measures for ethical use.

6.1 Study Overview

Our study aimed to develop a facial expression recognition model for intelligent surveillance systems using dynamic images, with a focus on accurately detecting emotions in real time to enhance security applications. The study involved comprehensive data acquisition, preprocessing, hyperparameter tuning, model selection, training, and evaluation. Evaluation metrics included accuracy, precision, recall, macro-averaged F1 score, and confusion matrices, along with computational measures such as system throughput and per-frame inference time to assess efficiency.

6.2 Main Contributions

This study advances facial expression recognition in surveillance systems through four main contributions:

1. **Design of a Deep Learning Dynamic Facial Expression Recognition Model:** We developed a CNN-LSTM hybrid architecture that combines MobileNetV2 for spatial feature extraction and LSTM layers for temporal modeling. The model achieves 95% accuracy, precision, recall, and F1-score on dynamic facial expression sequences, demonstrating its effectiveness in detecting critical emotions such as sadness and fear, where temporal dynamics play a key role in accurate recognition.

2. **Dynamic Evaluation of Facial Expression Sequences:** Unlike traditional static image approaches, this study focuses on dynamic facial sequences to capture temporal changes in expressions. Using the ADFES and Chinese Face Dataset, which include both face-forward and head-turning sequences, we thoroughly evaluated model performance across varying expression dynamics, head poses, and demographic groups. This approach establishes a more rigorous benchmark for dynamic facial expression recognition in surveillance settings.

3. **Ethical Deployment Guidelines:** Recognizing the privacy and security implications of emotion recognition in surveillance systems, we established comprehensive guidelines for responsible deployment. These guidelines address data governance, technical safeguards, transparency mechanisms, and regulatory compliance with frameworks such as POPIA, bridging the gap between technical capability and ethical implementation.

4. **Computational Real-Time Efficiency:** The proposed model demonstrated efficient real-time processing, achieving 43.09 FPS using sequence-based processing, a 3.09× improvement over single-frame processing (13.93 FPS) and exceeding the standard video requirement of 30 FPS. The reduced per-frame inference time from 0.0718 seconds to 0.0232 seconds makes this

approach computationally viable for live surveillance deployment while maintaining high classification accuracy.

6.3 Conclusion

Our study successfully developed and validated a CNN-LSTM hybrid model for facial expression recognition in intelligent surveillance systems. The model achieved 95% accuracy, precision, recall, and F1-score on dynamic facial expression sequences, demonstrating the effectiveness of combining MobileNetV2 for spatial feature extraction with LSTM layers for temporal modeling. The LSTM component proved particularly valuable for detecting emotions critical in surveillance scenarios, such as sadness and fear, where temporal dynamics provide essential discriminative information beyond static images.

Computational evaluation confirmed the feasibility of real-time deployment. Sequence-based processing achieved 43.09 FPS, exceeding the 30 FPS requirement for live video analysis and representing a 3.09× improvement over single-frame processing, with per-frame inference time reduced from 0.0718 seconds to 0.0232 seconds. This combination of high accuracy and computational efficiency demonstrates that the model can be deployed in operational surveillance systems without sacrificing performance or responsiveness.

However, practical deployment requires careful consideration of limitations and ethical implications. The model's performance declined under real-world conditions, such as occlusion and blur, highlighting the gap between controlled evaluation and operational surveillance environments. While our evaluation included diverse expression dynamics through the ADFES and Chinese Face Dataset, broader validation across demographic groups is essential to ensure equitable performance. These findings underscore that real-world deployment demands robust preprocessing, extensive data augmentation, diverse training data, and adherence to ethical guidelines to ensure reliable, fair, and responsible performance across populations and environmental conditions.

6.4 Limitations and Future Work

While our results demonstrate the potential of CNN-LSTM hybrid models for facial expression recognition, several limitations must be acknowledged. First, our evaluation was conducted on prerecorded datasets rather than in real-time implementation, which is critical for practical surveillance applications. Additionally, the model exhibits reduced robustness under real-world variations such as occlusion and blur, conditions commonly encountered in surveillance environments.

Future research can extend this study by testing the model in real-time surveillance scenarios to validate computational performance and reliability under live conditions. Incorporating adversarial training techniques, robust data augmentation, and domain adaptation methods could improve model resilience to environmental degradations frequently observed in surveillance settings. Moreover, including datasets representing diverse ethnicities and age groups would enhance generalizability, enabling more robust facial expression recognition across different demographics and reducing potential biases. Additionally, future studies should incorporate confidence intervals and statistical tests when comparing model performance, as this would provide more rigorous validation and substantiate observed performance differences between models. Furthermore, adopting cross-dataset evaluation protocols, such as training on one dataset and testing on an entirely different one, would offer a more rigorous assessment of model generalizability and robustness across varying demographic and cultural contexts.

Beyond generalizability, future work could explore optimization techniques such as quantization, pruning, and knowledge distillation to further compress the lightweight architecture. Such approaches would facilitate efficient deployment on edge devices, including embedded cameras and mobile surveillance systems, while maintaining high accuracy. These enhancements would increase the practical viability of real-time facial expression recognition in operational surveillance applications.

References

[1] Amina Adadi and Mohammed Berrada. "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)". In: *IEEE access* 6 (2018), pp. 52138–52160.

[2] S. Afra and R. Alhajj. "Early warning system: From face recognition by surveillance cameras to social media analysis to detecting suspicious people". In: *Physica A: Statistical Mechanics and its Applications* 540 (2020), p. 123151.

[3] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. "On the surprising behavior of distance metrics in high dimensional space". In: *International conference on database theory*. Springer. 2001, pp. 420–434.

[4] Charu C Aggarwal et al. *Neural networks and deep learning*. Vol. 10. 978. Springer, 2018.

[5] Muhammad Ahmed, Khurram Azeem Hashmi, Alain Pagani, Marcus Liwicki, Didier Stricker, and Muhammad Zeshan Afzal. "Survey and performance analysis of deep learning-based object detection in challenging environments". In: *Sensors* 21.15 (2021), p. 5116.

[6] I. Al-Amoudi, R. Samad, N. R. H. Abdullah, M. Mustafa, and D. Pebrianti. "Automatic attendance system using face recognition with deep learning algorithm". In: *Proc. 12th Nat. Tech. Seminar Unmanned Syst. Technol. 2020 (NUSYS'20)*. 2022, pp. 573–588. DOI: 10.1007/978-981-16-6471-3_42.

[7] Mona Ashok, Rohit Madan, Anton Joha, and Uthayasankar Sivarajah. "Ethi- cal framework for artificial intelligence and digital technologies". In: *Interna- tional Journal of Information Management* 62 (2022), p. 102433.

[8] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv". In: *arXiv preprint arXiv:1803.01271* 10 (2018).

[9] G. Bargshady, X. Zhou, R. C. Deo, J. Soar, F. Whittaker, and H. Wang. "En- hanced deep learning algorithm development to detect pain intensity from facial expression images". In: *Expert Syst. Appl.* 149 (2020), p. 113305. DOI: 10.1016/j.eswa.2020.113305.

[10] Lisa Feldman Barrett. "The theory of constructed emotion: an active inference account of interoception and categorization". In: *Social cognitive and affective neuroscience* 12.1 (2017), pp. 1–23.

[11] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. "Recognizing facial expression: machine learning and application to sponta- neous behavior". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE. 2005, pp. 568–573.

[12] D. Bhagat, A. Hanwate, R. V. Salunkhe, T. Jagyasi, P. Sardare, and M. Sahu. "Future Perspectives on Surveillance Systems". In: *Modern Advancements in Surveillance Systems and Technologies*. IGI Global Scientific Publishing, 2025,pp. 349–370.

[13] V. Bronstein. "Prioritising Command-and-Control Over Collaborative Gover- nance: The Role of the Information Regulator Under the Protection of Per- sonal Information Act". In: *Potchefstroom Electronic Law Journal (PELJ)* 25.1 (2022), pp. 1–41.

[14] A. Caroppo, A. Leone, and P. Siciliano. "Comparison between deep learn- ing models and traditional machine learning approaches for facial expression recognition in ageing adults". In: *Journal of Computer Science and Technology*35.5 (2020), pp. 1127–1146. DOI: 10.1007/s11390-020-9665-4.URL: https://doi.org/10.1007/s11390-020-9665-4.

[15] N. Choudhry, J. Abawajy, S. Huda, and I. Rao. "A comprehensive survey of machine learning methods for surveillance videos anomaly detection". In: *IEEE Access* 11 (2023), pp. 68616–68639. DOI: 10.1109/ACCESS.2023.3291566.

[16] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil. "Universals and cultural variations in 22 emotional expressions across five cultures". In: *Emotion* 18.1 (2018), pp. 75–93. DOI: 10.1037/emo0000302.

[17] N. Dalal and B. Triggs. "Histograms of oriented gradients for human detec- tion". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 886–893.

[18] P. Ekman and W. V. Friesen. "Constants across cultures in the face and emotion". In: *J. Personal. Soc. Psychol.* 17.2 (1971), pp. 124–129. DOI: 10 . 1037 / h0030377.

[19] O. S. Ekundayo and S. Viriri. "Facial expression recognition: A review of trends and techniques". In: *Ieee Access* 9 (2021), pp. 136944–136973.

[20] Y. Fan, X. Lu, D. Li, and Y. Liu. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks". In: *Proceedings of the 18th ACM In- ternational Conference on Multimodal Interaction*. 2016, pp. 445–450. DOI: 10 . 1145/2993148.2997632.

[21] H. Ghazouani. "Challenges and emerging trends for machine reading of the mind from facial expressions". In: *SN Computer Science* 5.1 (2023), p. 103.

[22] A. Gondkar, R. Gandhi, and N. Jadhav. "Facial emotion recognition using transfer learning: A comparative study". In: *2021 2nd global conference for advancement in technology (GCAT)*. IEEE. 2021, pp. 1–6.

[23] Muneeb Ul Hassan, Mubashir Husain Rehmani, and Jinjun Chen. "Differen- tial privacy techniques for cyber physical systems: A survey". In: *IEEE Communications Surveys & Tutorials* 22.1 (2019), pp. 746–789.

[24] A. G. Howard et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861* (2017).

[25] H. Huang, C. Qu, F. Xiang, and X. Li. "Facial Expression Recognition

UsingMobileNetV2 with Attention Mechanism and Facial Landmarks". In: (2023).H. Huang and C. Qu contributed equally to this work.

[26] S. A. Hussain and A. S. A. Al Balushi. "A real time face emotion classification and recognition using deep learning model". In: *J. Phys.: Conf. Ser.* Vol. 1432.1. 2020, p. 012087. DOI: 10.1088/1742-6596/1432/1/012087.

[27] B. Hutchinson et al. "Towards accountability for machine learning datasets: Practices from software engineering and infrastructure". In: *Proc. 2021 ACM Conf. Fairness, Accountability, Transparency*. 2021, pp. 560–575. DOI: 10.1145/3442188.3445918.

[28] Marijn Janssen, Paul Brous, Elsa Estevez, Luis S Barbosa, and Tomasz Janowski. "Data governance: Organizing data for trustworthy Artificial Intelligence". In: *Government information quarterly* 37.3 (2020), p. 101493.

[29] D. Jeong, B.-G. Kim, and S.-Y. Dong. "Deep joint spatiotemporal network (DJSTN) for efficient facial expression recognition". In: *Sensors* 20.7 (2020),p. 1936. DOI: 10.3390/s20071936.

[30] M. Karnati, A. Seal, D. Bhattacharjee, A. Yazidi, and O. Krejcar. "Understand- ing deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey". In: *IEEE Transactions on Instrumenttion and Measurement* 72 (2023), pp. 1–31.

[31] A. R. Khan. "Facial Emotion Recognition Using Conventional Machine Learn- ing and Deep Learning Methods: Current Achievements, Analysis and Re- maining Challenges". In: *Information* 13.268 (2022). DOI: 10.3390/info13060268. URL: https://doi.org/10.3390/info13060268.

[32] Mehran Khodabandeh. "Addressing the labeled data scarcity problem in deep learning". In: (2023).

[33] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras. "Advances in facial ex- pression recognition: A survey of methods, benchmarks, models, and datasets". In:

Information 15.3 (2024), p. 135. DOI: 10.3390/info15030135.

[34] M. Lades et al. "Distortion invariant object recognition in the dynamic link architecture". In: *IEEE Transactions on computers* 42.3 (1993), pp. 300–311.

[35] S. Li and W. Deng. "Deep facial expression recognition: A survey". In: *IEEE transactions on affective computing* 13.3 (2020), pp. 1195–1215.

[36] S. Li, W. Deng, and J. Du. "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[37] C. Lu et al. "Multiple spatio-temporal feature learning for video-based emo- tion recognition in the wild". In: *Proceedings of the 20th ACM International Con- ference on Multimodal Interaction*. 2018, pp. 646–652. DOI: 10.1145/3242969. 3243012.

[38] B. Martinez and M. F. Valstar. "Advances, challenges, and opportunities in automatic facial expression recognition". In: *Advances in face detection and facial image analysis* (2016), pp. 63–100.

[39] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopou- los. "Image segmentation using deep learning: A survey". In: *IEEE transac- tions on pattern analysis and machine intelligence* 44.7 (2021), pp. 3523–3542.

[40] A. Mollahosseini, D. Chan, and M. H. Mahoor. "Going deeper in facial expres- sion recognition using deep neural networks". In: *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE. 2016, pp. 1–10.

[41] A. Mollahosseini, B. Hasani, and M. H. Mahoor. "Affectnet: A database for fa- cial expression, valence, and arousal computing in the wild". In: *IEEE Trans- actions on Affective Computing* 10.1 (2017), pp. 18–31.

[42] M. Nazir, Z. Jan, and M. Sajjad. "Facial expression recognition using his- togram of oriented gradients based transformed features". In: *Cluster Com- puting* 21

(2018), pp. 539–548.

[43] I. Oztel, G. Yolcu, and C. Oz. “Performance comparison of transfer learn- ing and training from scratch approaches for deep facial expression recogni- tion”. In: *2019 4th International Conference on Computer Science and Engineering (UBMK)*. IEEE. 2019, pp. 1–6.

[44] Sinno Jialin Pan. “Transfer learning”. In: *Learning* 21 (2020), pp. 1–2.

[45] V. K. Patil, P. Nawade, R. Nagarkar, and P. Kadale. *Object Detection and Track- ing: Face Detection and Recognition*. Wiley Online Library, 2024, pp. 25–54.

[46] A. A. Pise, M. A. Alqahtani, P. Verma, D. A. Karras, A. Halifa, et al. “Methods for facial expression recognition with applications in challenging situations”. In: *Comput. Intell. Neurosci.* 2022 (2022), pp. 1–18. DOI: 10.1155/2022/9214065.

[47] D. Rangulov and M. Fahim. “Emotion recognition on large video dataset based on convolutional feature extractor and recurrent neural network”. In: *2020 IEEE 4th International Conference on Image Processing, Applications and Sys- tems (IPAS)*. 2020, pp. 14–20. DOI: 10.1109/IPAS50080.2020.9334976.

[48] A. Rehman, M. Mujahid, A. Elyassih, B. AlGhofaily, and S. A. O. Bahaj. “Comprehensive Review and Analysis on Facial Emotion Recognition: Performance Insights into Deep and Traditional Learning with Current Updates and Challenges”. In: *Computers, Materials & Continua* 82.1 (2025).

[49] J. A. Russell. “Facial expressions of emotion: What lies beyond minimal universality?” In: *Psychological Bulletin*. Vol. 118. 3. 1995, pp. 379–391. DOI: 10.1037/0033-2909.118.3.379.

[50] James A Russell. “Is there universal recognition of emotion from facial expres- sion? A review of the cross-cultural studies.” In: *Psychological bulletin* 115.1 (1994), p. 102.

[51] James A Russell. “Core affect and the psychological construction of emotion.” In:

Psychological review 110.1 (2003), p. 145.

[52] I. Saadi, T.-A. Abdelmalik, A. Hadid, Y. El Hillali, et al. "Driver's facial ex- pression recognition: A comprehensive survey". In: *Expert Syst. Appl.* (2023),p. 122784. DOI: 10.1016/j.eswa.2023.122784.

[53] F. Z. Salmam, A. Madani, and M. Kissi. "Facial expression recognition using decision trees". In: *2016 13th international conference on computer graphics, imag- ing and visualization (CGiV)*. IEEE. 2016, pp. 125–130.

[54] M. Shafiq and Z. Gu. "Deep residual learning for image recognition: A sur- vey". In: *Appl. Sci.* 12.18 (2022), p. 8972. DOI: 10.3390/app12188972.

[55] B. Shahriari, A. Bouchard-Côté, and N. Freitas. "Unbounded Bayesian op- timization via regularization". In: *Artificial intelligence and statistics*. PMLR. 2016, pp. 1168–1176.

[56] C. Shan, S. Gong, and P. W. McOwan. "Facial expression recognition based on local binary patterns: A comprehensive study". In: *Image and vision Computing* 27.6 (2009), pp. 803–816.

[57] C. Shorten and T. M. Khoshgoftaar. "A survey on image data augmentation for deep learning". In: *J. Big Data* 6.1 (2019), pp. 1–48. DOI: 10.1186/s40537- 019-0197-0.

[58] A. Singh, S. Bhatt, V. Nayak, and M. Shah. "Automation of surveillance sys- tems using deep learning and facial recognition". In: *Int. J. Syst. Assur. Eng. Manag.* 14 (2023), pp. 236–245. DOI: 10.1007/s13198-022-01751-w.

[59] D. Sinha and M. El-Sharkawy. "Thin mobilenet: An enhanced mobilenet ar- chitecture". In: *2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*. IEEE. 2019, pp. 0280–0285.

[60] J. Snoek, H. Larochelle, and R. P. Adams. "Practical bayesian optimization of machine learning algorithms". In: *Advances in neural information processing systems*

25 (2012).

[61] A. S. M. Sohail and P. Bhattacharya. “Classification of facial expressions using k-nearest neighbor classifier”. In: *Computer Vision/Computer Graphics Collabo- ration Techniques: Third International Conference, MIRAGE 2007, Rocquencourt, France, March 28-30, 2007. Proceedings 3*. Springer. 2007, pp. 555–566.

[62] G. Sreenu and M. A. Saleem Durai. “Intelligent video surveillance: a review through deep learning techniques for crowd analysis”. In: *Journal of Big Data*6.1 (2019),p.48.DOI: URL:https://doi.org/10.1186/s40537-019-0212-5.

[63] G. Sunkara. “Ethical and regulatory implications of AI in cybersecurity surveillance”. In: *World Journal of Advanced Research and Reviews* 24.2 (2024), pp. 2895–2905. DOI: 10.30574/wjarr.2024.24.2.3571.

[64] M. Tan and Q. Le. “Efficientnet: Rethinking model scaling for convolutional neural networks”. In: *International conference on machine learning*. PMLR. 2019,pp. 6105–6114.

[65] S. Tariq, M. Baruwal Chhetri, S. Nepal, and C. Paris. “Alert fatigue in security operations centres: Research challenges and opportunities”. In: *ACM Comput- ing Surveys* 57.9 (2025), pp. 1–38.

[66] Mohammad Mustafa Taye. “Understanding of machine learning with deep learning: architectures, workflow, applications and future directions”. In: *Computers* 12.5 (2023), p. 91.

[67] Ehsan Toreini, Maryam Mehrnezhad, and Aad van Moorsel. “Fairness as a Service (FaaS): verifiable and privacy-preserving fairness auditing of machine learning systems”. In: *International Journal of Information Security* 23.2 (2024),

pp. 981–997.

[68] T. Valeeprakhon, K. Orkphol, and P. Chaihuadjaroen. “Deep convolutional neural networks based on VGG-16 transfer learning for abnormalities peeled

shrimp classification". In: *Int. Sci. J. Eng. Technol.* 6.2 (2022), pp. 13–23.

[69] M. R. Venkateshwarlu, N. S. Bhargavi, T. V. Kumar, K. S. Kumar, P. Jash- wanth, S. K. Mudhiraj, et al. "Facial Expression Recognition System Using Deep Learning Models Based on Human Emotions through Classification with CNN, RNN, and YOLO Object Detection Algorithms". In: *International Journal of Information Technology and Computer Engineering* 12.2 (2024), pp. 401– 406.

[70] S. Vosta and K.-C. Yow. "A CNN-RNN combined structure for real-world violence detection in surveillance cameras". In: *Appl. Sci.* 12.3 (2022), p. 1021.DOI: 10.3390/app12031021.

[71] Youzi Xiao et al. "A review of object detection based on deep learning". In: *Multimedia Tools and Applications* 79.33 (2020), pp. 23729–23791.

[72] Runhua Xu, Nathalie Baracaldo, and James Joshi. "Privacy-preserving machine learning: Methods, challenges and directions". In: *arXiv preprint arXiv:2108.04417*(2021).

[73] Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. "How fast are the leaked facial expressions: The duration of micro-expressions". In: *Journal of nonverbal behavior* 37.4 (2013), pp. 217–230.

[74] Y. Yin, M. Ayoub, A. Abel, and H. Zhang. "A dynamic emotion recognition system based on convolutional feature extraction and recurrent neural network". In: *Intell. Syst. Appl., Proc. 2022 Intell. Syst. Conf. (IntelliSys)*. Vol. 2. 2022, pp. 134–154. DOI: 10.1007/978-3-031-16078-3_8.

[75] G. Yolcu et al. "Facial expression recognition for monitoring neurological dis- orders based on convolutional neural network". In: *Multimedia Tools Appl.* 78 (2019), pp. 31581–31603. DOI: 10.1007/s11042-019-07865-x.

[76] K. Zhang, Y. Huang, Y. Du, and L. Wang. "Facial expression recognition based on deep evolutional spatial-temporal networks". In: *IEEE Transactions on Image*

Processing 26.9 (2017), pp. 4193–4203.

[77] X. Zhang, X. Zhou, M. Lin, and J. Sun. "Shufflenet: An extremely efficient convolutional neural network for mobile devices". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6848–6856.

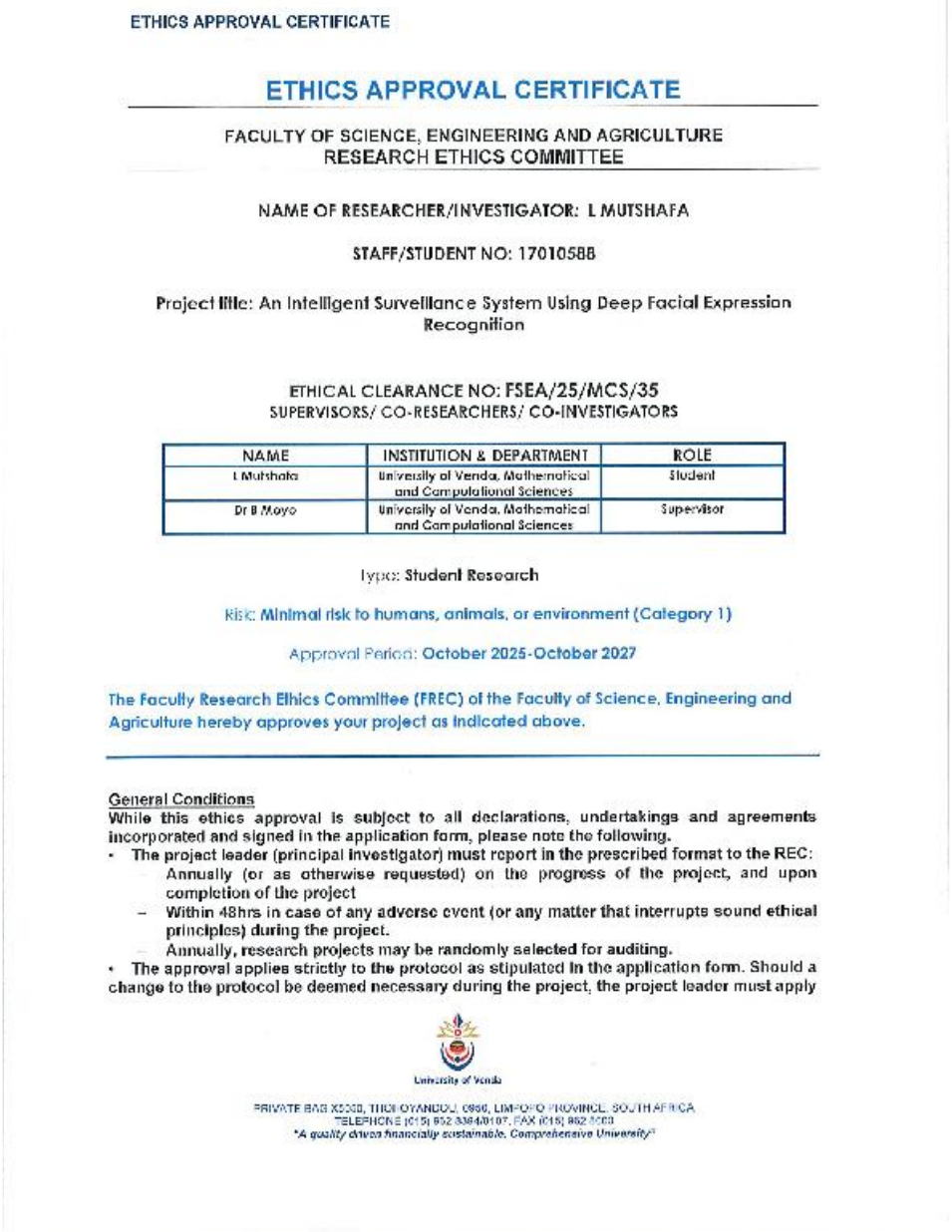
[78] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. "Object detection with deep learning: A review". In: *IEEE transactions on neural networks and learning systems* 30.11 (2019), pp. 3212–3232.

[79] N. Zhou, R. Liang, and W. Shi. "A lightweight convolutional neural network for real-time facial expression detection". In: *IEEE Access* 9 (2020), pp. 5573– 5584.

[80] M. Zhuang et al. "Highly robust and wearable facial expression recognition via deep-learning-assisted, soft epidermal electronics". In: *Research* 2021 (2021),p. 9761736. DOI: 10.34133/2021/9761736.

[81] Matsumoto, D., Hwang, H.C. Clusters of Nonverbal Behaviors Differ According to Type of Question and Veracity in Investigative Interviews in a Mock Crime Context. *J Police Crim Psych* **33**, 302–315 (2018). https://doi.org/10.1007/s11896-017-9250-0

APPENDIX A: ETHICAL CLEARANCE



ETHICS APPROVAL CERTIFICATE

ETHICS APPROVAL CERTIFICATE

FACULTY OF SCIENCE, ENGINEERING AND AGRICULTURE
RESEARCH ETHICS COMMITTEE

NAME OF RESEARCHER/INVESTIGATOR: L MUTSHAFA

STAFF/STUDENT NO: 17010588

Project title: An Intelligent Surveillance System Using Deep Facial Expression Recognition

ETHICAL CLEARANCE NO: FSEA/25/MCS/35
SUPERVISORS/ CO-RESEARCHERS/ CO-INVESTIGATORS

NAME	INSTITUTION & DEPARTMENT	ROLE
L Mutshafa	University of Venda, Mathematical and Computational Sciences	Student
Dr B Moyo	University of Venda, Mathematical and Computational Sciences	Supervisor

Type: Student Research

Risk: Minimal risk to humans, animals, or environment (Category 1)

Approval Period: October 2025-October 2027

The Faculty Research Ethics Committee (FREC) of the Faculty of Science, Engineering and Agriculture hereby approves your project as indicated above.

General Conditions
While this ethics approval is subject to all declarations, undertakings and agreements incorporated and signed in the application form, please note the following.

- The project leader (principal investigator) must report in the prescribed format to the REC:
  - Annually (or as otherwise requested) on the progress of the project, and upon completion of the project
  - Within 48hrs in case of any adverse event (or any matter that interrupts sound ethical principles) during the project.
  - Annually, research projects may be randomly selected for auditing.
- The approval applies strictly to the protocol as stipulated in the application form. Should a change to the protocol be deemed necessary during the project, the project leader must apply

University of Venda

PRIVATE BAG X5050, THOHOYANDOU, 0950, LIMPOPO PROVINCE, SOUTH AFRICA
TELEPHONE (015) 962-8384/0107, FAX (015) 962-2003
"A quality driven financially sustainable, Comprehensive University"

LETTER

ETHICS APPROVAL CERTIFICATE

for approval of these changes before their implementation. Should there be a deviation from the study protocol, without the necessary approval for the change, the ethics approval is automatically forfeited.

- The date of approval indicates the earliest date that the project may begin. Should the project have to continue after the expiry date; a new application must be made, and a new approval received before or on the expiry date.
- In the interest of ethical responsibility, the FREC retains the right to:
  - Request access to any information or data at any time during the course or after completion of the project,
  - To ask further questions; Seek additional information; Require further modification or monitor the conduct of your research or the informed consent process.
  - withdraw or postpone approval if:
  - Any unethical principles or practices of the project are revealed or suspected.
  - It becomes apparent that relevant information was withheld from the REC or that information has been false or misrepresented.
  - The required annual report and reporting of adverse events was not done timely and accurately,
  - New institutional rules, national legislation or international conventions deem it necessary

---

ISSUED BY:
FACULTY OF SCIENCE, ENGINEERING AND AGRICULTURE RESEARCH ETHICS COMMITTEE
Date considered: 15 October 2025

Chairperson: Prof LC Murulana

Signature: .....

FACULTY OF SCIENCE,
ENGINEERING AND AGRICULTURE
OFFICE OF THE EXECUTIVE DEAN
DEPUTY DEAN : RESEARCH AND POST GRADUATE STUDIES
UNIVERSITY OF VENDA

DATE: .....
NAME: ..... SIGNATURE: .....
This document is approved for further processing

University of Venda

PRIVATE BAG X5050, THOHOYANDOU, 0950, LIMPOPO PROVINCE, SOUTH AFRICA
TELEPHONE (015) 962 8304/5107, FAX (015) 962 8000
"A quality driven financially sustainable, Comprehensive University"