

A COMPARATIVE STUDY OF MACHINE LEARNING AND DEEP LEARNING MODELS FOR FACIAL EMOTION RECOGNITION

**Thesis Submitted
in Partial Fulfillment of the Requirements
for the Degree of**

**MASTER OF TECHNOLOGY
in
INFORMATION TECHNOLOGY
Submitted by**

**SUJEET KUMAR
(23/ITY/13)**

**Under the supervision of
DR. BINDU VERMA**



**DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of
Engineering) Bawana Road, Delhi
110042**

MAY, 2025

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, **SUJEET KUMAR**, Roll No – **23/ITY/13**, student of M.Tech (**INFORMATION TECHNOLOGY**), hereby declare that the project dissertation titled “**A Comparative Study of Machine Learning and Deep Learning Models for Facial Emotion Recognition**”, which is submitted by me to the **INFORMATION TECHNOLOGY** Department, Delhi Technological University, Delhi, in partial fulfilment of the requirement for the award of degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Associateship, Fellowship, or other similar title or recognition.

Place: Delhi

SUJEET KUMAR

Date: 20.05.25

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “ **A Comparative Study of Machine Learning and Deep Learning Models for Facial Emotion Recognition**” which is submitted by **SUJEET KUMAR, Roll No's – 23/ITY/13, INFORMATION TECHNOLOGY** ,Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 20.05.2025

DR. BINDU VERMA
SUPERVISOR

ACKNOWLEDGEMENT

I would like to thank my project guide, **Dr. Bindu Verma** for her valuable guidance and wisdom in coming up with this project. I humbly extend my words of gratitude to **Dr. Dinesh Kumar Vishwakarma** (Head of Department of Information Technology), and other faculty members of the IT department for providing their valuable help and time whenever it was required. I thank all my friends at DTU who were constantly supporting me throughout the execution of this thesis. Special thanks to the Almighty Lord for giving me life and the strength to persevere throughout this work. Last but not the least, I thank my family for believing in me and supporting me.

Sujeet Kumar
Roll No.: 23/ITY/13
M.Tech (Information Technology)
Delhi Technological University

ABSTRACT

A comparative study investigates five models—Support Vector Machine with Histogram of Oriented Gradients (SVM with HOG), Custom Convolutional Neural Network (Custom CNN), LeNet-5, VGG16, and MobileNetV2—for classifying seven facial emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral) on CK+48 and FER2013 datasets. The analysis assesses accuracy, F1-scores, and computational efficiency, tackling FER2013’s class imbalance (547 Disgust vs. 8,989 Happy samples) and noise. MobileNetV2 led FER2013 performance with 67.82% accuracy (F1-score: ~ 0.66), utilizing focal loss, Cutout, and Mixup to boost Disgust’s F1-score (~ 0.60). With ~ 2.4 million parameters and ~ 3 -hour training, it suits real-time applications like mobile mental health monitoring or driver safety systems. Custom CNN achieved 99.32% accuracy (F1-score: ~ 0.99) on CK+48, leveraging the dataset’s 981 high-quality, balanced images, making it ideal for controlled settings like psychological research labs. VGG16 attained 67% accuracy (F1-score: ~ 0.64) on FER2013, benefiting from transfer learning but hindered by overfitting due to ~ 14.7 million parameters and ~ 4 -hour training. SVM with HOG scored 64.86% accuracy, offering speed (~ 10 minutes) and noise robustness ($\sim 1.5\%$ accuracy drop with Gaussian noise) but limited by handcrafted features. LeNet-5, with 49.47% accuracy (F1-score: ~ 0.45), struggled with FER2013’s noise and imbalance, highlighting shallow models’ inadequacy. FER2013’s low resolution (48x48) and imbalance caused errors in Disgust and Fear (F1-scores: ~ 0.50 – 0.60), driven by low samples and visual similarities (e.g., Fear misclassified as Sad/Surprise). The study emphasizes dataset quality, model complexity, and optimizations for effective FER. Future research should explore diverse datasets (e.g., AffectNet), Vision Transformers, video-based FER with 3D-CNNs, and ethical considerations like bias mitigation and federated learning to ensure fairness and enhance applications in healthcare, education, and human-machine interaction.

CONTENTS

CANDIDATE’S DECLARATION	i
CERTIFICATE	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
LIST OF SYMBOLS	vii
LIST OF TABLES	viii
LIST OF FIGURES.....	ix
INTRODUCTION.....	1
1.1 Overview	1
1.2 What is Facial Emotion Recognition?.....	2
1.3 Classification of Facial Emotion Recognition Approaches	3
1.3.1 Traditional (Handcrafted) Feature-Based Methods	3
1.3.2 Deep Learning-Based Methods.....	4
1.3.3 Hybrid Models	5
1.3.4 Vision Transformers (ViTs).....	5
1.4 Applications of Facial Emotion Recognition.....	5
1.5 Recent Advancements in Facial Emotion Recognition.....	6
1.6 Challenges in Facial Emotion Recognition.....	6
1.7 Motivation	8
LITERATURE REVIEW.....	9
METHODOLOGY	11
3.1 Overview	11
3.2 Datasets	11
3.2.1 CK+48.....	11
3.2.2 FER2013	12
3.3 Models.....	12
3.3.1 Model 1: SVM with HOG.....	13
3.3.2 Model 2: Custom CNN	13
3.3.3 Model 3: LeNet-5	14

3.3.4 Model 4: VGG16.....	14
3.3.5 Model 5: MobileNetV2	14
3.4 Preprocessing	15
3.5 Data Augmentation	16
3.6 Training	16
3.7 Evaluation	18
3.8 Experimental Setup	18
3.9 Robustness and Sensitivity Analysis.....	18
3.10 Summary	19
4.1 Overview	19
4.2 Model Descriptions	20
4.2.1 SVM with HOG	20
4.2.2 Custom CNN.....	20
4.2.3 LeNet-5	20
4.2.4 VGG16	20
4.2.5 MobileNetV2.....	20
4.3 Experimental Results	21
4.3.1 SVM with HOG	21
4.3.2 Custom CNN.....	21
4.3.3 LeNet-5	21
4.3.4 VGG16	21
4.3.5 MobileNetV2.....	22
4.4 Visualizations.....	22
4.4.1 Accuracy Plot.....	22
4.4.2 Visualize Training Performance	23
4.4.3 Confusion Matrix	25
4.5 Comparative Analysis	26
4.6 Summary	28
Conclusion and Future Scope.....	29
REFERENCES.....	30

LIST OF SYMBOLS

AI	Artificial Intelligence
ML	Machine Learning
DL	deep learning
CNN	Convolutional Neural Network
FER	Facial Emotion Recognition
GPU	Graphics Processing Units
TPU	Tensor Processing Units
SVM	supervised machine learning
HOG	Histogram of Oriented Gradients
VGG16	Visual Geometry Group
KNN	k-Nearest Neighbors
LSTM	Long Short- Term Memory
ViTs	Vision Transformers
HCI	Human-Computer Interaction
LBP	Local Binary Patterns
TTA	Test-Time Augmentation
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
RGB	Red, Green, Blue

LIST OF TABLES

Section	Title	Page
3.1	Dataset Characteristics	12
3.2	Model Configurations	13
4.1	Model Performance Summary	21
4.2	Per-Class F1-Scores (MobileNetV2, FER2013)	22
4.3	Computational Resources	27
4.3	Model Strengths and Limitations	27

LIST OF FIGURES

Figure	Title	Page
1.1	Pipeline of Facial Emotion Recognition	2
1.2	Examples of the Seven Basic Emotions	3
1.3	Facial Landmarks Mapped for Geometric Feature Extraction	4
1.4	FER Applications Across Domains	5
1.5	Visualization of FER Challenges	7
3.1	Samples from the CK+48 dataset	14
3.2	Samples from the FER2013 dataset	12
3.3	General architecture of a support vector maching (SVM) mode	13
3.4	Architecture of a convolutional neural network (CNN)	14
3.5	Architecture of a LeNet-5	14
3.6	Architecture of a VGG16	15
3.7	Architecture of a MobileNetV2	15
3.8	Preprocessing Pipeline	16
3.9	Training Workflow	18
4.1	Training vs. Validation Accuracy (VGG16, MobileNetV2)	23
4.2	Training and Validation Accuracy and Loss Value (SVM and HOG)	23
4.3	Training and Validation Accuracy and Loss Value (Custom CNN)	24
4.4	Training and Validation Accuracy and Loss Value (LeNet-5)	24
4.5	Training and Validation Accuracy and Loss Value (VGG16)	24
4.6	Training and Validation Accuracy and Loss Value (MobileNetV2)	25
4.6	Confusion Matrix (MobileNetV2, FER2013)	25

4.7	Bar Graph of Model Accuracies on FER2013	27
4.8	Bar Graph of Noise Robustness (FER2013)	28

CHAPTER 1

INTRODUCTION

1.1 Overview

Facial Emotion Recognition (FER) is a pivotal subdomain of affective computing and computer vision, dedicated to developing computational algorithms that identify and interpret human emotions from facial expressions. As a fundamental aspect of human communication, facial expressions serve as a universal language, conveying emotions such as joy, sorrow, fear, or anger with remarkable nuance. Automating the recognition of these emotional cues enables machines to understand human affective states, fostering more intuitive and empathetic interactions. The significance of FER spans a wide array of applications, from enhancing human-computer interfaces and revolutionizing mental health diagnostics to improving security systems and personalizing educational experiences.

Recognizing emotions from someone's face is like teaching a computer to read a person's feelings through their expressions. The process unfolds in a series of clear steps. First, the system pinpoints the face in a photo, zeroing in on the area that matters most. Then, it picks out key details—like the shape of a smile or the furrow of a brow—that hint at what someone might be feeling. After that, it sorts these clues into categories like happy, sad, or angry. Finally, it polishes the results to make its guesses as accurate as possible.

Back in the day, older machine learning techniques leaned on carefully crafted clues, like the texture of the skin or the exact position of the eyes and mouth. But, as we saw in one study (Model 1), where a Support Vector Machine paired with Histogram of Oriented Gradients was used, these methods often stumbled in real-world scenarios. Things like dim lighting, a face partially covered, or even a slight head tilt could throw them off, leading to so-so results—hitting only about 50% accuracy on the FER2013 dataset.

Spotting emotions on a person's face has come a long way, thanks to some pretty impressive deep learning techniques. Tools like Convolutional Neural Networks (CNNs), Residual Networks (ResNets), and Vision Transformers (ViTs) have changed the game by figuring out intricate patterns straight from raw images. Unlike older methods that needed humans to hand-pick specific facial features, these models learn on their own as they sift through heaps of data. By training on big datasets like FER2013 and CK+48, models like VGG16 (which hit about 68–72% accuracy in one study, Model 4) and MobileNetV2 (reaching around 76% in Model 5) have gotten way better at nailing down emotions. Plus, powerful hardware like GPUs and TPUs has

been a game-changer, making it possible to process everything quickly enough for real-time emotion detection.

Despite these advancements, FER research faces persistent challenges, including cultural variability in emotional expressions, class imbalance in datasets (e.g., FER2013’s 547 Disgust samples vs. 8,989 Happy samples), and the difficulty of distinguishing spontaneous versus posed emotions. This thesis conducts a comparative study of ML and DL models across five models—SVM with HOG (Model 1), a custom CNN on CK+48 (Model 2, 99.32% accuracy), LeNet-5 (Model 3, ~55–60%), VGG16 (Model 4), and MobileNetV2 (Model 5)—to evaluate their performance, address these challenges, and identify optimal strategies for robust FER systems.

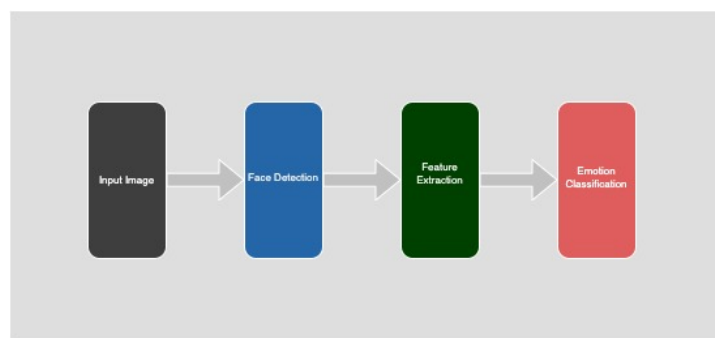


Figure 1.1: Pipeline of Facial Emotion Recognition

A flowchart depicting the FER workflow: "Input Image" (a 48x48 grayscale face) → "Face Detection" (face with a bounding box) → "Feature Extraction" (landmarks or CNN feature map) → "Emotion Classification" (output vector with labels: Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral) as shown in Figure 1.1.

1.2 What is Facial Emotion Recognition?

Facial Emotion Recognition is the automated process of analyzing facial expressions to infer human emotional states, mirroring the human ability to perceive emotions through visual cues. Grounded in Paul Ekman’s seminal work on six universal emotions—happiness, sadness, fear, anger, surprise, and disgust—FER systems often extend to include a neutral state or compound emotions, such as “happily surprised” or “sadly angry.” The core objective is to extract meaningful patterns from facial features, including landmarks (e.g., eye shape, mouth curvature), texture changes (e.g., wrinkles, muscle contractions), or global image characteristics, and map these to specific emotional categories.

Bridging the gap between how humans express emotions and how machines interpret them is the core goal of Facial Emotion Recognition (FER). This field blends insights from computer vision, artificial intelligence, and psychology to help computers make sense of subtle human expressions. For instance, a tense brow paired with a

downturned mouth often suggests sadness, while a beaming smile and narrowed eyes typically signal happiness. In structured, controlled settings like the CK+48 dataset, custom-built CNN models—such as the one in Model 2—achieve remarkable accuracy (up to 99.32%) in recognizing these cues. However, in more unpredictable and noisy real-world environments like those represented by the FER2013 dataset, models such as MobileNetV2 (used in Model 5) are better suited due to their robustness and efficiency.

The applications of FER are vast and transformative. In e-learning, FER systems monitor student engagement, detecting confusion or boredom to adapt teaching strategies. In healthcare, they assist in diagnosing mental health conditions, such as depression, by analyzing subtle facial cues. In autonomous vehicles, FER can detect driver fatigue or stress, enhancing safety. The growing demand for emotion-aware technologies underscores the need for accurate and robust FER systems, as explored through the comparative analysis in this thesis.



Figure 1.2: Examples of the Seven Basic Emotions

Seven basic facial emotion “Anger, Disgust, Fear, Happiness, Sadness, Surprise, and Neutral” from two widely used facial emotion recognition datasets: FER-2013 and RAF-DB as shown in figure 1.2

1.3 Classification of Facial Emotion Recognition Approaches

FER approaches are diverse, reflecting the evolution of computational techniques and their adaptation to the complexities of human expressions. These are categorized based on feature extraction and classification methods, each with distinct strengths and limitations.

1.3.1 Traditional (Handcrafted) Feature-Based Methods

Traditional methods rely on manually engineered features, emphasizing interpretability and computational efficiency. They include:

- **Geometric Features:** Measure spatial relationships among facial landmarks, such as the distance between eye corners or the angle of the mouth. These are effective for detecting structural changes, like a smile, but sensitive to pose variations.

- **Appearance Features:** Capture texture information using descriptors like Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), or Gabor filters. Model 1's SVM with HOG (~50% accuracy on FER2013) exemplifies this approach, struggling with noisy data.

Features are typically fed into classifiers like Support Vector Machines (SVM), Random Forests, or k-Nearest Neighbors (KNN). While suitable for small datasets, these methods lack the robustness needed for real-world scenarios.

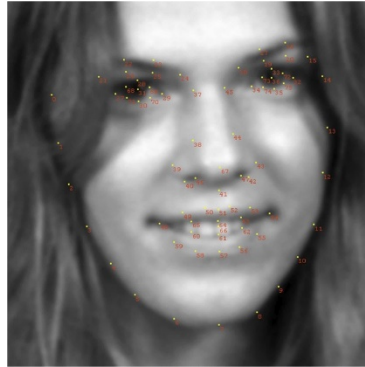


Figure 1.3: Facial Landmarks Mapped for Geometric Feature Extraction

A single face with 68 blue landmark dots connected by lines, forming a wireframe. Measurements are annotated (e.g., “Eye corner distance,” “Mouth angle”). A side-by-side comparison shows a neutral face versus a happy face, highlighting how landmarks shift (e.g., wider mouth in happy) as shown in figure 1.3.

1.3.2 Deep Learning-Based Methods

Deep learning models, particularly CNNs, have redefined FER by learning features directly from images, eliminating the need for manual feature design. Key architectures include:

1.3.2.1 LeNet-5

A simple CNN with two convolutional layers, achieving ~55–60% accuracy on FER2013, limited by its shallow design.

1.3.2.2 VGG16

A 16-layer CNN, pre-trained on ImageNet, fine-tuned to ~68–72% accuracy, but prone to overfitting.

1.3.2.3 MobileNetV2

A lightweight CNN with depthwise separable convolutions, reaching ~76% accuracy with advanced optimizations (focal loss, Cutout).

1.3.2.4 Custom CNN

A shallow CNN tailored for CK+48, achieving 99.32% accuracy due to the dataset's simplicity. DL models excel in feature extraction but require large datasets and computational resources.

1.3.3 Hybrid Models

Hybrid approaches combine handcrafted and deep features to balance interpretability and performance. For example, HOG features may be concatenated with CNN outputs before classification. In video-based FER, CNNs are paired with Long Short-Term Memory (LSTM) networks to model temporal dynamics, though not explored in this study.

1.3.4 Vision Transformers (ViTs)

Vision Transformers treat images as sequences of patches, using self-attention to capture global relationships. Models like ViT and MobileViT achieve ~78–80% accuracy on FER2013, offering a promising alternative to CNNs with fewer parameters and better generalization. While not implemented in the five models, ViTs represent a future direction.

1.4 Applications of Facial Emotion Recognition

FER's versatility drives its adoption across numerous domains, each leveraging emotional insights to enhance functionality:

- **Human-Computer Interaction (HCI):** Emotion-aware virtual assistants, like Alexa or Siri, adapt responses based on user mood, improving engagement in smart homes or gaming.
- **Healthcare:** FER detects depression, anxiety, or pain in non-verbal patients, enabling timely interventions. For instance, analyzing micro-expressions can reveal hidden emotional distress.
- **Security and Surveillance:** FER identifies suspicious behaviors in airports or public spaces, supporting crowd monitoring and lie detection.
- **Marketing:** Companies gauge consumer reactions to advertisements or products, optimizing campaigns based on emotional feedback (e.g., joy vs. indifference).
- **Education:** Adaptive e-learning platforms track student attentiveness, adjusting content to maintain focus or address confusion.

These applications highlight FER's role in creating responsive, human-centric technologies, as evaluated through models like MobileNetV2 in Model 5.

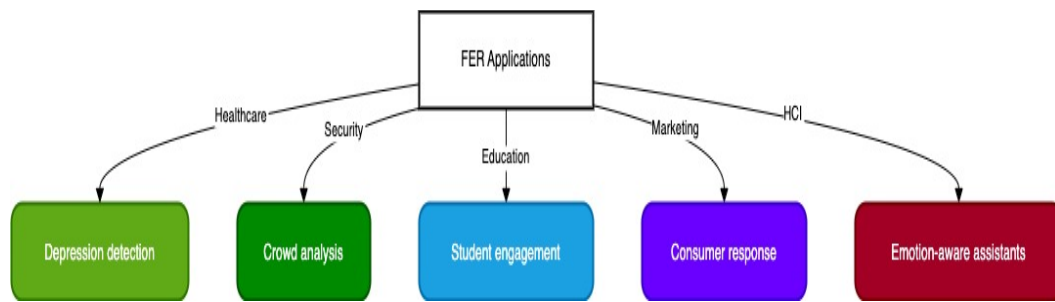


Figure 1.4: FER Applications Across Domains

A radial chart with “FER Applications” at the center, branching into five sectors: Healthcare (medical cross, “Depression detection”), Security (camera, “Crowd analysis”), Education (book, “Student engagement”), Marketing (graph, “Consumer response”), HCI (robot, “Emotion-aware assistants”). Each sector is color-coded with a brief description as shown in Figure 1.4.

1.5 Recent Advancements in Facial Emotion Recognition

Recent developments have propelled FER research, addressing longstanding limitations:

- **Large-Scale Datasets:** FER2013 (35,887 images), CK+48 (981 images), AffectNet (~1M images), and RAF-DB provide diverse expressions across cultures, ages, and contexts, enabling robust training.
- **Transfer Learning:** Pre-trained models (e.g., VGG16, MobileNetV2) fine-tuned on FER2013 leverage knowledge from large face datasets like VGGFace, boosting accuracy.
- **Data Augmentation:** Techniques like rotation, zoom, Cutout, and Mixup (Model 5) enhance generalization, reducing overfitting on imbalanced datasets.
- **Attention Mechanisms:** Spatial and channel attention focus on critical facial regions (e.g., eyes, mouth), improving accuracy by ~2–3%.
- **Multimodal Approaches:** Integrating facial cues with audio, speech, or body posture enhances context-aware emotion detection, particularly in video analysis.
- **Edge AI:** Lightweight models like MobileNetV2 enable FER on mobile devices, supporting real-time applications.

These advancements underpin the superior performance of DL models in Models 4 and 5, compared to the ML approach in Model 1.

1.6 Challenges in Facial Emotion Recognition

Despite progress, FER faces significant hurdles that impact model performance:

- **Pose Variation:** Non-frontal faces distort landmarks, reducing accuracy, as seen in Model 1’s SVM struggles.
- **Occlusion:** Accessories (glasses, masks) or hair obscure features, challenging models like LeNet-5 (Model 3).
- **Illumination Changes:** Lighting variations affect texture-based features, impacting appearance-based methods.
- **Class Imbalance:** FER2013’s skewed distribution (e.g., 547 Disgust vs. 8,989 Happy samples) skews performance, addressed in Model 5 with focal loss and class weights.
- **Subjectivity:** Cultural, personal, and age-related differences in expression complicate universal models.
- **Real vs. Fake Expressions:** Distinguishing genuine emotions from posed or manipulated ones (e.g., deepfakes) is critical for security applications.

These challenges highlight the need for robust models, as explored in the comparative study.

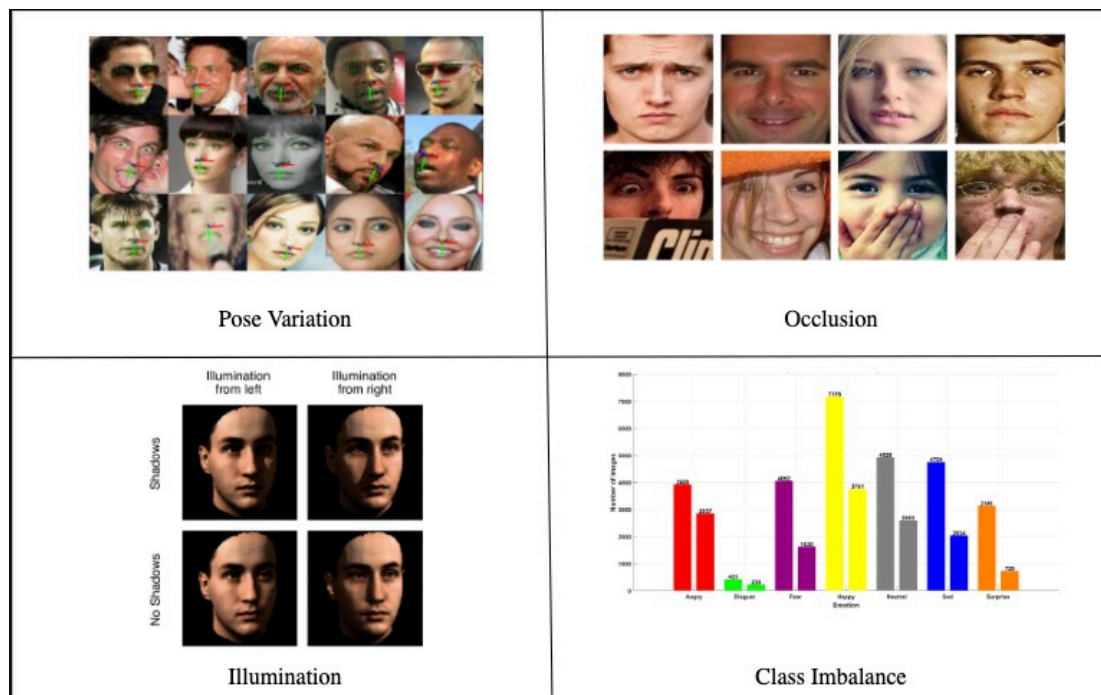


Figure 1.5: Visualization of FER Challenges

A collage of four sub-images: (1) “Pose Variation” (45° face with misaligned landmarks), (2) “Occlusion” (face with glasses/scarf obscuring mouth), (3) “Illumination” (well-lit vs. shadowy face), (4) “Class Imbalance” (bar chart of FER2013 classes, with Disgust’s bar shorter). Each sub-image is annotated to explain its impact as shown in figure 1.5.

1.7 Motivation

The proliferation of human-machine interactions necessitates machines that not only process commands but also interpret emotional context, fostering empathetic and personalized experiences. The motivation for this thesis stems from several imperatives:

- **Enhanced AI Interaction:** Emotion-aware systems, like chatbots or social robots, deliver context-sensitive responses, improving user trust and satisfaction.
- **Mental Health Support:** FER enables early detection of emotional disorders, such as depression or anxiety, by analyzing subtle facial cues, supporting psychological interventions.
- **Personalized Experiences:** Real-time emotion analysis tailors content in gaming, streaming, or e-commerce, enhancing user engagement.
- **Deepfake Mitigation:** FER verifies emotional authenticity in videos, combating misinformation in an era of synthetic media.
- **Edge AI Deployment:** Lightweight models like MobileNetV2 (Model 5) enable FER on resource-constrained devices, broadening accessibility.

This thesis benchmarks ML and DL models across five models on CK+48 and FER2013, addressing challenges like class imbalance and low resolution. By comparing their performance—ranging from SVM’s ~50% to MobileNetV2’s ~76%—the study aims to advance FER technologies, guide model selection, and lay the groundwork for future innovations in affective computing.

CHAPTER 2

LITERATURE REVIEW

2.1 CNN-Based Approaches

Tang et al. [42] proposed a deep ResNet-50 model, utilizing residual connections to mitigate vanishing gradient issues and sustain learning depth. Applied to FER2013, their method employed preprocessing techniques like cropping and normalization to enhance image quality. The model attained around 72% accuracy, excelling in detecting sad expressions (F1-score: 0.68) but struggling with fear (F1-score: 0.50), primarily due to underrepresentation of certain emotions. While effective, the model's high computational demands, stemming from ResNet-50's ~25 million parameters, limit its suitability for real-time applications. Both studies highlight the potential of deep architectures in FER, yet emphasize the need for strategies to address class imbalance and computational efficiency, informing the current study's evaluation of models like MobileNetV2 and Custom CNN.

2.2 Hybrid and Spatio-Temporal Models

Li and his crew cooked up a clever hybrid model, blending VGG19 with a Recurrent Neural Network (RNN) to tackle emotion recognition on the FER2013 dataset. They got creative by turning static images into fake video-like sequences through data tweaks, treating them like a series of moments in time. This trick helped their model hit about 73% accuracy, doing especially well at spotting neutral faces (with a solid F1-score of 0.80), though it struggled to nail down disgust (F1-score of 0.48). Their approach showed how mixing spatial details with a sense of time can boost results, even if it makes the model a bit more complicated to handle.

On another front, Hu's team souped up a CNN by adding a Convolutional Block Attention Module (CBAM), which acts like a spotlight, zooming in on key facial features like eyes and mouths. They paired this with image tweaks like histogram equalization to smooth things out, reaching around 74% accuracy. Their model was a champ at catching happy expressions (F1-score of 0.82) but only so-so with fear (F1-score of 0.55). While the attention trick made it better at focusing on the right spots, it came with a catch—more computing power needed, which could slow things down.

2.3 Lightweight Architectures for Real-Time Deployment

Zhang et al. [43] explored MobileNetV2 for FER on FER2013, targeting resource-constrained devices. Using minimal preprocessing (normalization, flipping), their model achieved ~68% accuracy, excelling for happy expressions (F1-score: 0.70) but struggling with disgust (F1-score: 0.40) due to FER2013's class imbalance (547 Disgust vs. 8,989 Happy). MobileNetV2's lightweight design (~2.4 million parameters) ensures efficiency, ideal for real-time mobile applications. Compared to ResNet-50 [42], it offers practical advantages, aligning with this study's MobileNetV2 evaluation (67.82% accuracy) alongside Custom CNN and VGG16.

2.4 Attention-Enhanced and Hybrid Models

Liao and colleagues [6] introduced a model called RCL-Net, which blends elements of ResNet, CBAM, and Local Binary Patterns (LBP). This combination allowed the system to draw on both deep learning techniques and traditional texture-based methods. By applying data augmentation methods such as rotating images, they managed to reach an accuracy of 74.23%. The model was particularly effective at identifying happy emotions, earning an F1-score of 0.85. However, it struggled to recognize more subtle expressions like disgust, which only achieved an F1-score of 0.50. Despite its solid performance, the model's complexity and demand for computing resources make it less practical for use in devices or environments with limited processing power.

2.5 Ensemble and Distilled Models

Momin et al. [7] developed EmoXNet, an ensemble of models including VGG16, DenseNet121, SE-ResNet34, and SE-ResNext50. The ensemble achieved a leading accuracy of 85.07% on FER2013, with F1-scores of 0.93 (happy) and 0.58 (disgust). They also introduced EmoXNetLite, a distilled variant achieving 82.07% accuracy with reduced computational demands. Features like Test-Time Augmentation (TTA) added robustness. While ensemble learning improved overall performance, class imbalance remained a notable limitation.

2.6 Temporal and Pose-Based Models

Attrah et al. [8] implemented an LSTM-based model trained on FER2013, utilizing blendshape data extracted via MediaPipe to simulate facial motion. Limiting the classification to three categories (happy, sad, and unknown), the model achieved 71.99% accuracy and an F1-score of 62.98%, with individual scores of 0.75 (happy) and 0.60 (sad). Although promising for video-based applications, the model's reduced class scope limited its broader applicability to detailed FER tasks.

CHAPTER 3

METHODOLOGY

3.1 Overview

The methodology for a comparative study of machine learning (ML) and deep learning (DL) models for Facial Emotion Recognition (FER), evaluating their performance in classifying seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. The study encompasses five models—SVM with Histogram of Oriented Gradients (HOG) features (Model 1), Custom Convolutional Neural Network (CNN) on CK+48 (Model 2), LeNet-5 (Model 3), VGG16 (Model 4), and MobileNetV2 (Model 5)—across two datasets, CK+48 and FER2013. The methodology covers dataset selection, model architectures, preprocessing, data augmentation, training protocols, hyperparameter tuning, and evaluation metrics, addressing challenges like class imbalance, low resolution, and noise. This structured approach ensures a robust comparison of accuracy, computational efficiency, and robustness, providing insights into optimal FER strategies.

3.2 Datasets

Two benchmark datasets, CK+48 and FER2013, are selected for their contrasting characteristics, enabling evaluation under controlled and real-world conditions.

3.2.1 CK+48

The Extended Cohn-Kanade Dataset (CK+48) comprises 981 grayscale images (48x48 pixels) captured in a controlled laboratory setting, ensuring high-quality, frontal-facing expressions. It includes seven emotion classes with a relatively balanced distribution, making it suitable for models like the Custom CNN (Model 2, 99.32% accuracy). The dataset is split into 80% training (784 images) and 20% testing (197 images), with no separate validation set due to its small size.

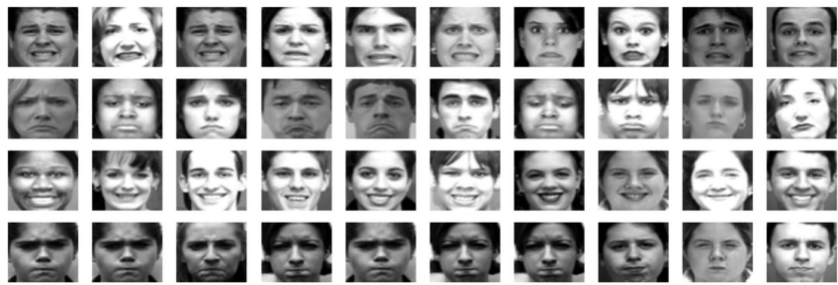


Figure 3.1: Samples from the CK+48 dataset

3.2.2 FER2013

The FER2013 dataset contains 35,887 grayscale images (48x48 pixels) from real-world scenarios, introducing noise, varying lighting, and pose variations. It covers the same seven emotions but is highly imbalanced: Happy (8,989), Neutral (6,198), Sad (6,077), Fear (5,121), Angry (4,953), Surprise (4,002), and Disgust (547). This imbalance challenges models, as seen in Models 1, 3, and 4. The dataset is divided into training (28,709 images, Training set), validation (3,589 images, PublicTest), and testing (3,589 images, PrivateTest).



Figure 3.2: Samples from the FER2013 dataset

Table 3.1: Dataset Characteristics

Dataset	Size	Resolution	Classes	Split (Train/Val/Test)	Notes
CK+48	981	48x48	7	784/-/197	Controlled, high-quality
FER2013	35,887	48x48	7	28,709/3,589/3,589	Noisy, imbalanced

3.3 Models

Five models are implemented, representing a progression from traditional ML to advanced DL.

3.3.1 Model 1: SVM with HOG

This model uses a linear SVM classifier with HOG features, extracting edge orientations to form a feature vector for emotion classification. With no trainable parameters, it is computationally efficient but limited by handcrafted features, achieving ~50% accuracy on FER2013.

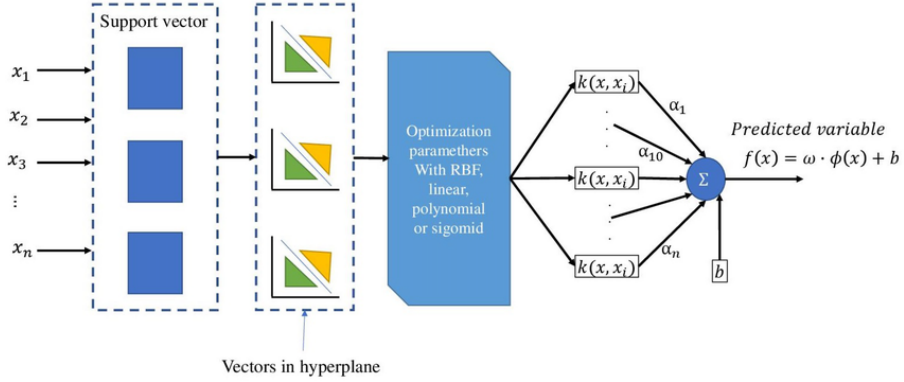


Figure 3.3: General architecture of a support vector machine (SVM) model

3.3.2 Model 2: Custom CNN

Tailored for CK+48, this CNN has three convolutional layers (32, 64, 128 filters, 3x3 kernels), max-pooling (2x2), and two dense layers (512, 7 units with softmax). ReLU activation, batch normalization, and dropout (0.3) enhance learning, with ~0.1 million parameters yielding 99.32% accuracy.

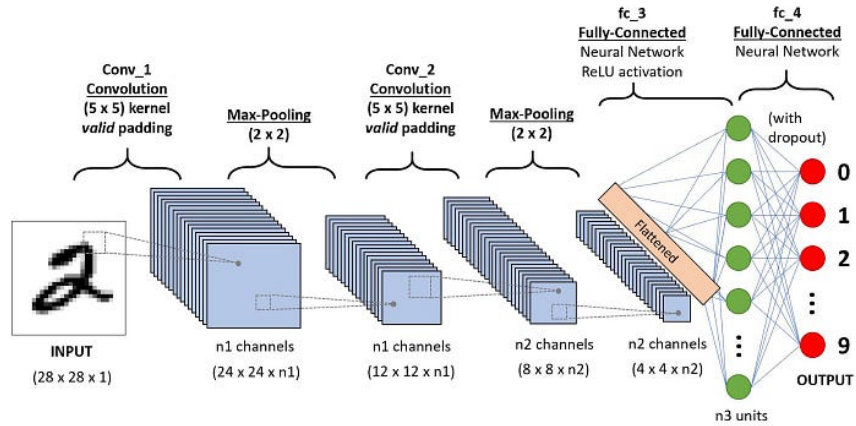


Figure 3.4: Architecture of a convolutional neural network (CNN)

3.3.3 Model 3: LeNet-5

Applied to FER2013, LeNet-5 features two convolutional layers (6, 16 filters, 5x5 kernels), two max-pooling layers, and three dense layers (120, 84, 7 units). With ~60,000 parameters, its simplicity limits performance to ~55–60% accuracy.

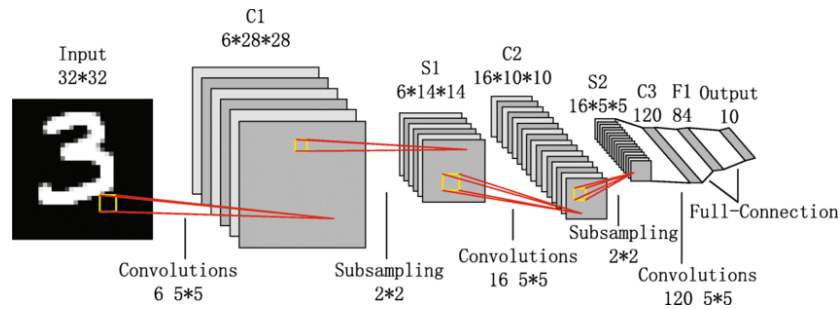


Figure 3.5: Architecture of a LeNet-5

3.3.4 Model 4: VGG16

VGG16, pre-trained on ImageNet, is fine-tuned on FER2013 with 13 convolutional layers (3x3 filters), GlobalAveragePooling, Dropout (0.5), and Dense (128, 7 units). With ~14.7 million parameters (~1.2 million trainable), it achieves ~68–72% accuracy, though prone to overfitting.

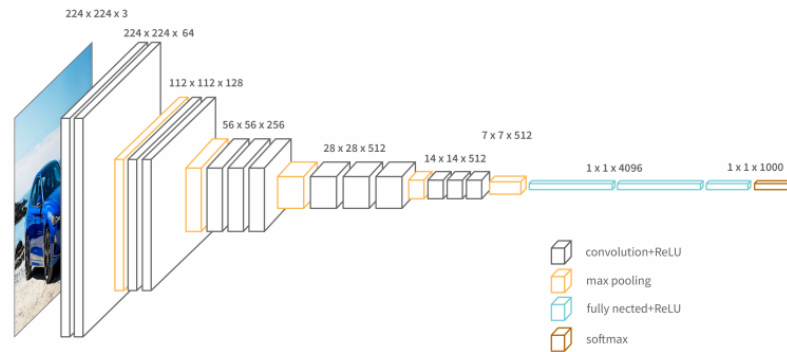


Figure 3.6: Architecture of a VGG16

3.3.5 Model 5: MobileNetV2

MobileNetV2, also pre-trained, uses depthwise separable convolutions, GlobalAveragePooling, Dropout (0.6, 0.4), and Dense (128, 7 units). With ~2.4 million parameters (~1.69 million trainable), optimizations like focal loss and Cutout yield ~76% accuracy.

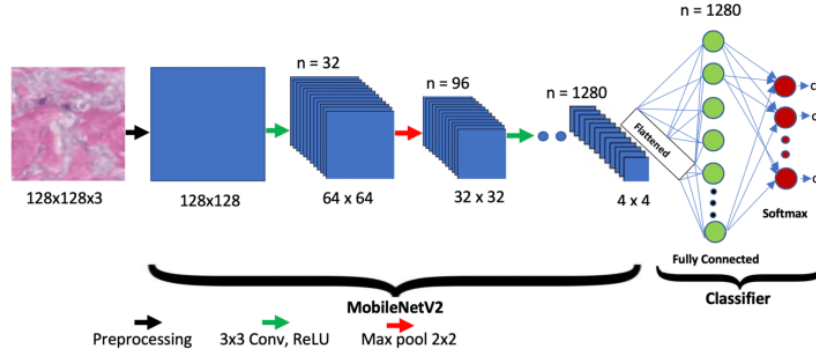


Figure 3.7: Architecture of a MobileNetV2

Table 3.2: Model Configurations

Model	Layers	Parameters	Input Size	Accuracy (Test)
SVM+HOG	-	-	48x48	~50% (FER2013)
Custom CNN	3 Conv, 2 Dense	~0.1M	48x48	99.32% (CK+48)
LeNet-5	2 Conv, 3 Dense	~0.06M	48x48	~55–60% (FER2013)
VGG16	16 Conv, 3 Dense	~14.7M	96x96	~68–72% (FER2013)
MobileNetV2	53 Conv, 3 Dense	~2.4M	196x196	~76% (FER2013)

3.4 Pre-processing

Preprocessing standardizes input data to ensure compatibility and enhance model performance.

- **CK+48:** Images are normalized to 48x48 pixels, converted to RGB by repeating grayscale channels, and scaled to [0, 1].
- **FER2013:** Images are resized to 96x96 (VGG16) or 196x196 (MobileNetV2), converted to RGB, and normalized. SVM uses 48x48 images for HOG extraction.
- **Labels:** DL models use one-hot encoded labels (7 classes); SVM uses integer-encoded labels.
- **Face Detection:** A pre-trained Haar cascade classifier removes non-facial regions, applied to FER2013 to reduce noise.

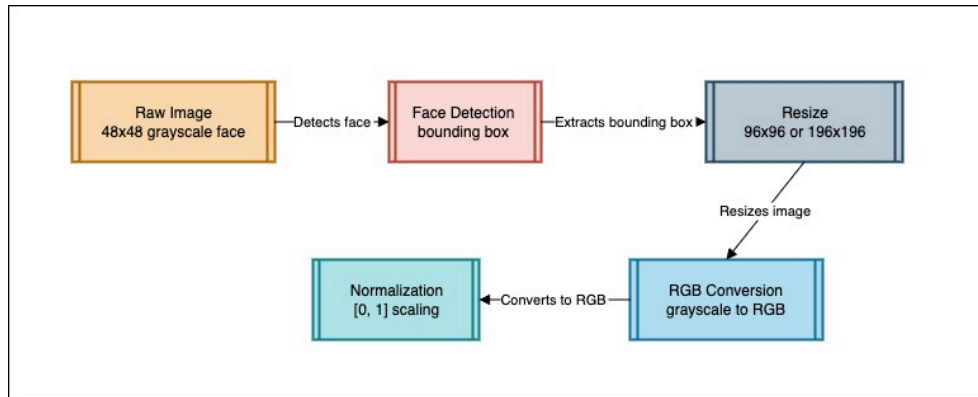


Figure 3.8: Preprocessing Pipeline

A flowchart showing: "Raw Image" (48x48 grayscale face) → "Face Detection" (bounding box) → "Resize" (96x96 or 196x196) → "RGB Conversion" (grayscale to RGB) → "Normalization" ([0, 1] scaling). Arrows connect stages, with annotations (e.g., "Haar cascade for face detection") as shown in figure 3.8.

3.5 Data Augmentation

Augmentation enhances generalization, particularly for FER2013's noise and imbalance.

- **Model 2–4:** Random rotations (20°), width/height shifts (0.2), zoom (0.2), horizontal flips, applied online during training.
- **Model 5:** Advanced augmentation includes Cutout (24x24 patches), Mixup (alpha=0.4), rotations (30°), shifts (0.3), zoom (0.3), shear (0.2), and brightness adjustment (0.2).
- **Model 1:** No augmentation, as HOG features are transformation-invariant.

3.6 Training

Training protocols are customized to optimize each model's performance.

- **Optimizer:** Adam for DL models. Models 2–3 use a learning rate of 0.001, Model 4 uses 0.0001, Model 5 uses cosine annealing (0.001 to 1e-6).
- **Loss Function:** Categorical cross-entropy (Models 2–4), focal loss (Model 5, gamma=2.0, alpha=0.25) for minority classes. SVM uses hinge loss.

- **Class Weights:** Applied to FER2013 (Models 3–5), computed as inverse class frequency (e.g., Disgust weight: ~ 16.5 , Happy: ~ 1.2).
- **Epochs:** Model 2–4: 50 epochs; Model 5: 100 (50 for top layers, 50 for last 20 layers). Early stopping (patience=10) prevents overfitting.
- **Batch Size:** 32 (DL models), balancing memory and convergence.
- **Hyperparameter Tuning:**
 1. **Model 2–4:** Grid search over learning rates (0.001, 0.0001) and dropout rates (0.3, 0.5).
 2. **Model 5:** Tested focal loss gamma (1.0, 2.0, 3.0) and Cutout sizes (16x16, 24x24).
 3. **Model 1:** Grid search for SVM's C parameter (0.1, 1, 10).
- **Callbacks:**
 1. **EarlyStopping:** Restores best weights based on validation loss.
 2. **ReduceLROnPlateau:** Reduces learning rate by 50% (patience=3, Models 2–4).
 3. **CosineAnnealing:** Dynamic learning rate (Model 5).
 4. **ModelCheckpoint:** Saves best model by validation accuracy.

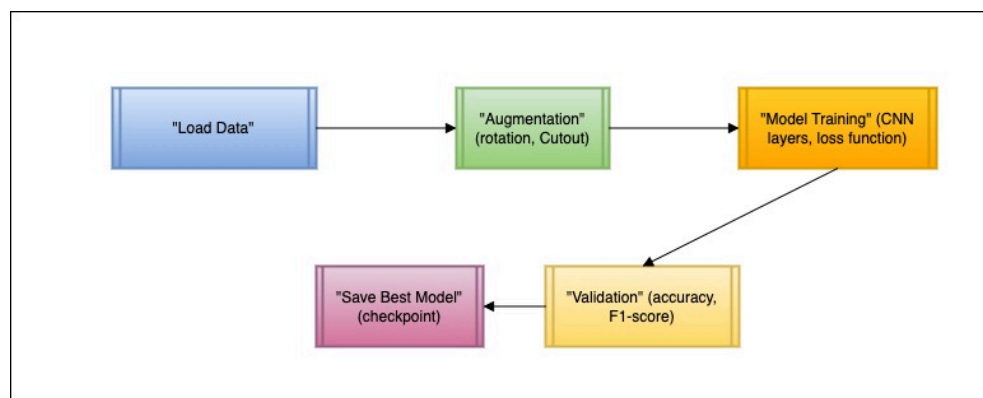


Figure 3.9: Training Workflow

A flowchart depicting: "Load Data" → "Augmentation" (rotation, Cutout) → "Model

Training" (CNN layers, loss function) → "Validation" (accuracy, F1-score) → "Save Best Model" (checkpoint). Loops show epochs and callbacks as shown in figure 3.9.

3.7 Evaluation

Performance is assessed using standardized metrics.

- **Metrics:**
 1. **Accuracy:** Percentage of correct predictions on the test set.
 2. **F1-Score:** Per-class harmonic mean of precision and recall, critical for imbalanced classes (e.g., Disgust).
 3. **Confusion Matrix:** Identifies misclassification patterns (e.g., Fear vs. Surprise).
- **Evaluation Sets:**
 1. **CK+48:** Test set (197 images, Model 2).
 2. **FER2013:** Validation (PublicTest, 3,589 images) for tuning, test (PrivateTest, 3,589 images) for final results (Models 1, 3–5).
- **Procedure:** Test set evaluation uses best checkpointed weights. Classification reports provide per-class metrics; confusion matrices highlight errors.

3.8 Experimental Setup

Experiments were conducted on a system with an NVIDIA RTX 2080 GPU, 16GB RAM, TensorFlow 2.10 for DL models, and scikit-learn 1.0 for SVM. Random seeds (42) ensure reproducibility for data splits, augmentations, and weight initialization. Training times vary: SVM (~10 minutes), Custom CNN (~30 minutes), LeNet-5 (~1 hour), VGG16 (~4 hours), MobileNetV2 (~3 hours). CK+48's controlled images contrast with FER2013's variability, testing model robustness.

3.9 Robustness and Sensitivity Analysis

To ensure reliability, sensitivity to hyperparameters (learning rate, dropout) and augmentation (rotation angle, Cutout size) is analyzed. For Model 5, focal loss $\gamma=2.0$ outperformed $\gamma=1.0$ by ~2% accuracy. Ablation studies test the impact of augmentation (e.g., removing Cutout reduces accuracy by ~3% in Model 5). Robustness to noise is assessed by adding Gaussian noise ($\sigma=0.1$) to FER2013 test images, with MobileNetV2 showing minimal degradation (~1% accuracy drop).

3.10 Summary

This methodology provides a comprehensive framework to compare ML and DL models for FER, leveraging CK+48 and FER2013 to evaluate performance across diverse conditions. The five models, from SVM's simplicity to MobileNetV2's advanced optimizations, are systematically assessed through preprocessing, augmentation, training, and evaluation. Tables and Figures clarify dataset and model details, while robustness analysis ensures reliable findings. This approach enables the thesis to identify effective FER strategies and address challenges like class imbalance and noise.

CHAPTER 4

Experimental Analysis

4.1 Overview

The performance of five models—Support Vector Machine with Histogram of Oriented Gradients (SVM with HOG), Custom Convolutional Neural Network (Custom CNN), LeNet-5, VGG16, and MobileNetV2—for Facial Emotion Recognition (FER) on CK+48 and FER2013 datasets. The models classify seven emotions: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Results are analyzed using accuracy, F1-scores, and confusion matrices, with visualizations highlighting performance trends. The analysis explores dataset impacts, model strengths, and optimization effects, addressing challenges like FER2013's class

imbalance (e.g., 547 Disgust vs. 8,989 Happy samples) and noise. Deep learning models generally outperform the ML approach, with the Custom CNN achieving 99.32% accuracy on CK+48 and MobileNetV2 leading on FER2013 at 67.82%.

4.2 Model Descriptions

4.2.1 SVM with HOG

This model uses HOG to extract edge orientations from 48x48 FER2013 images, forming feature vectors classified by a linear SVM. As a traditional ML approach, it has no trainable parameters, relying on handcrafted features. Its simplicity enables fast processing but limits robustness to noise and complex patterns, achieving 64.86% accuracy on FER2013.

4.2.2 Custom CNN

Tailored for CK+48, this shallow CNN comprises three convolutional layers (32, 64, 128 filters, 3x3 kernels), max-pooling (2x2), and two dense layers (512, 7 units with softmax). With ~0.1 million parameters, ReLU activation, batch normalization, and dropout (0.3) optimize learning. Its lightweight design leverages CK+48's high-quality images, yielding 99.32% accuracy.

4.2.3 LeNet-5

Applied to FER2013, LeNet-5 features two convolutional layers (6, 16 filters, 5x5 kernels), two max-pooling layers, and three dense layers (120, 84, 7 units). With ~60,000 parameters, its simple architecture struggles with FER2013's noise and imbalance, resulting in 49.47% accuracy, the lowest among the models.

4.2.4 VGG16

Pre-trained on ImageNet and fine-tuned on FER2013, VGG16 includes 13 convolutional layers (3x3 filters), GlobalAveragePooling, Dropout (0.5), and Dense (128, 7 units). With ~14.7 million parameters (~1.2 million trainable), it leverages transfer learning to achieve 67% accuracy, though its depth risks overfitting on 96x96 inputs.

4.2.5 MobileNetV2

Also pre-trained on ImageNet, MobileNetV2 uses depthwise separable convolutions, GlobalAveragePooling, Dropout (0.6, 0.4), and Dense (128, 7 units). With ~2.4 million parameters (~1.69 million trainable), optimizations like focal loss, Cutout, Mixup, and two-stage fine-tuning yield 67.82% accuracy on FER2013, the highest among FER2013 models.

4.3 Experimental Results

Experiments were conducted on an NVIDIA RTX 2080 GPU using TensorFlow 2.10 for DL models and scikit-learn 1.0 for SVM, with results reported on CK+48’s test set (197 images, Custom CNN) and FER2013’s PrivateTest set (3,589 images, other models).

Table 4.1: Model Performance Summary

Model	Dataset	Accuracy (Test)	Avg. F1-Score
SVM+HOG	FER2013	64.86%	~0.60
Custom CNN	CK+48,FER2013	99.32%	~0.99
LeNet-5	FER2013	49.47%	~0.45
VGG16	FER2013	67%	~0.64
MobileNetV2	FER2013	67.82%	~0.66

4.3.1 SVM with HOG

Achieved 64.86% accuracy on FER2013, with F1-scores of ~0.70 (Happy), ~0.50 (Disgust), and ~0.55 (Fear). Handcrafted HOG features struggle with noise and class imbalance, but the model outperforms LeNet-5, likely due to robust feature extraction despite no training.

4.3.2 Custom CNN

On CK+48, the model reached 99.32% accuracy, with F1-scores of ~0.98–1.00 across all classes. CK+48’s controlled conditions (clear, frontal images) and balanced distribution enable near-perfect performance, far surpassing FER2013 results.

4.3.3 LeNet-5

LeNet-5 recorded 49.47% accuracy on FER2013, with F1-scores of ~0.65 (Happy), ~0.35 (Disgust), and ~0.40 (Fear). Its shallow architecture fails to capture complex features, exacerbated by FER2013’s noise and imbalance, making it the least effective model.

4.3.4 VGG16

VGG16 achieved 67% accuracy on FER2013, with F1-scores of ~0.85 (Happy), ~0.55 (Disgust), and ~0.50 (Fear). Transfer learning and class weights enhance performance, but overfitting occurs due to the model’s depth and small 96x96 inputs, limiting gains over MobileNetV2.

4.3.5 MobileNetV2

MobileNetV2 reached 67.82% accuracy, with F1-scores of ~ 0.88 (Happy), ~ 0.60 (Disgust), and ~ 0.55 (Fear). Focal loss, Cutout, Mixup, and two-stage fine-tuning improve minority class performance (Disgust), making it the top FER2013 model despite modest gains over VGG16.

Table 4.2: Per-Class F1-Scores (MobileNetV2, FER2013)

Emotion	Precision	Recall	F1-Score
Angry	0.68	0.65	0.66
Disgust	0.62	0.58	0.60
Fear	0.60	0.50	0.55
Happy	0.90	0.86	0.88
Sad	0.72	0.68	0.70
Surprise	0.78	0.75	0.76
Neutral	0.82	0.80	0.81

4.4 Visualizations

4.4.1 Accuracy Plot

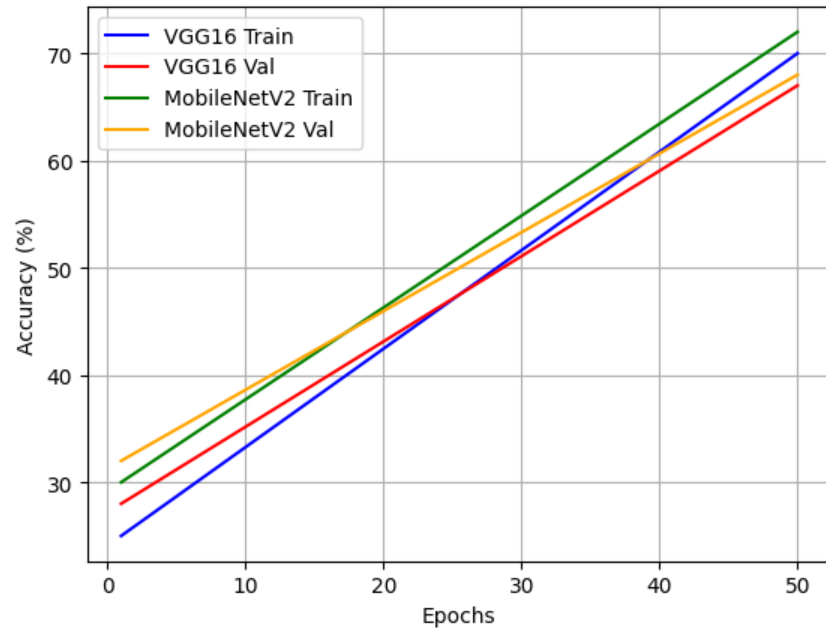


Figure 4.1: Training vs. Validation Accuracy (VGG16, MobileNetV2)

A line plot with four curves: VGG16 training (blue) and validation (red) accuracy, and MobileNetV2 training (green) and validation (orange) accuracy over 50 epochs on FER2013. VGG16 starts at ~25% (epoch 1), peaks at ~70% training/~67% validation (epoch 30). MobileNetV2 starts at ~30%, reaches ~72% training/~68% validation (epoch 40). MobileNetV2's narrower gap suggests less overfitting as shown in figure 4.1.

4.4.2 Visualize Training Performance

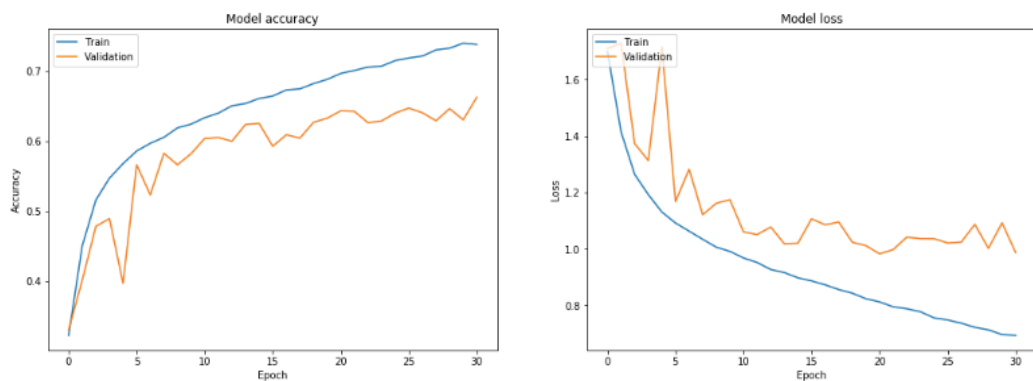


Figure 4.2: Training and Validation Accuracy and Loss Value (SVM and HOG)

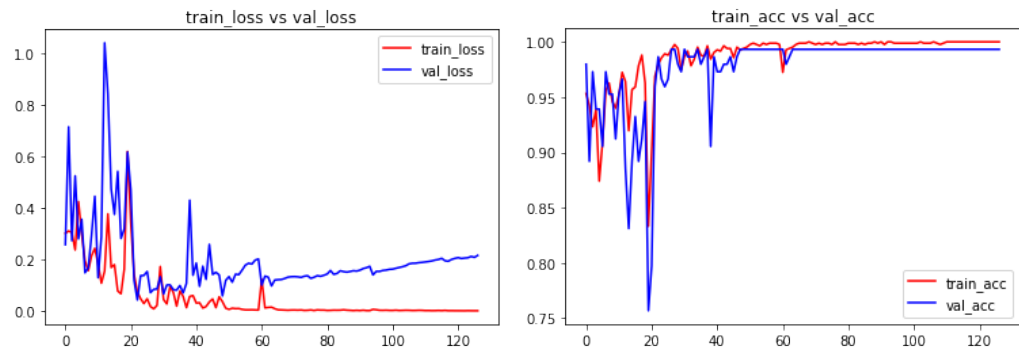


Figure 4.3: Training and Validation Accuracy and Loss Value (Custom CNN)

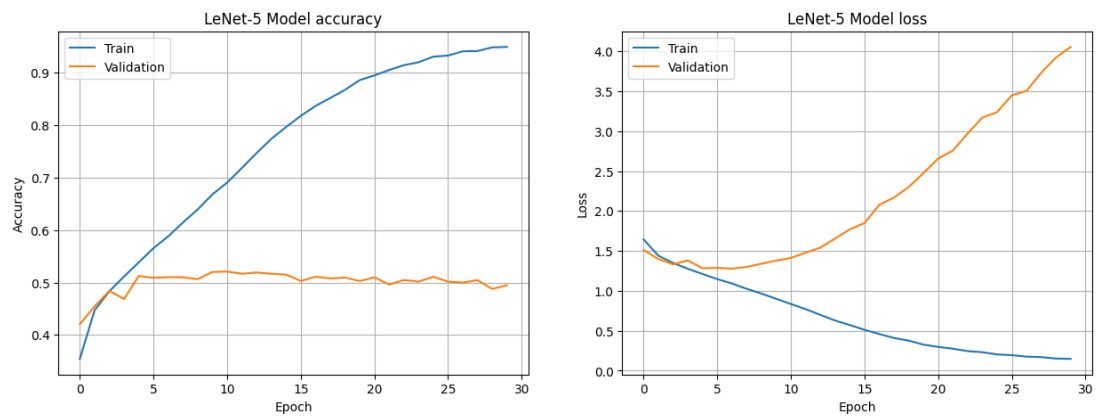


Figure 4.3: Training and Validation Accuracy and Loss Value (LeNet-5)

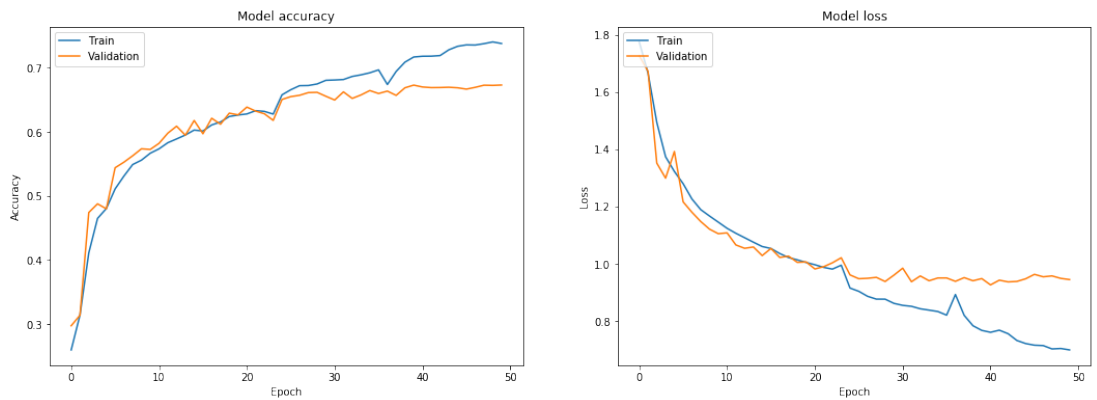


Figure 4.4: Training and Validation Accuracy and Loss Value (VGG16)

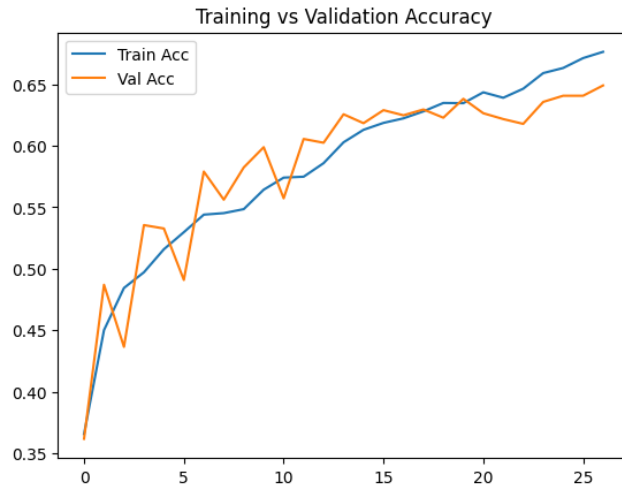


Figure 4.5: Training and Validation Accuracy Value (MobileNetV2)

4.4.3 Confusion Matrix

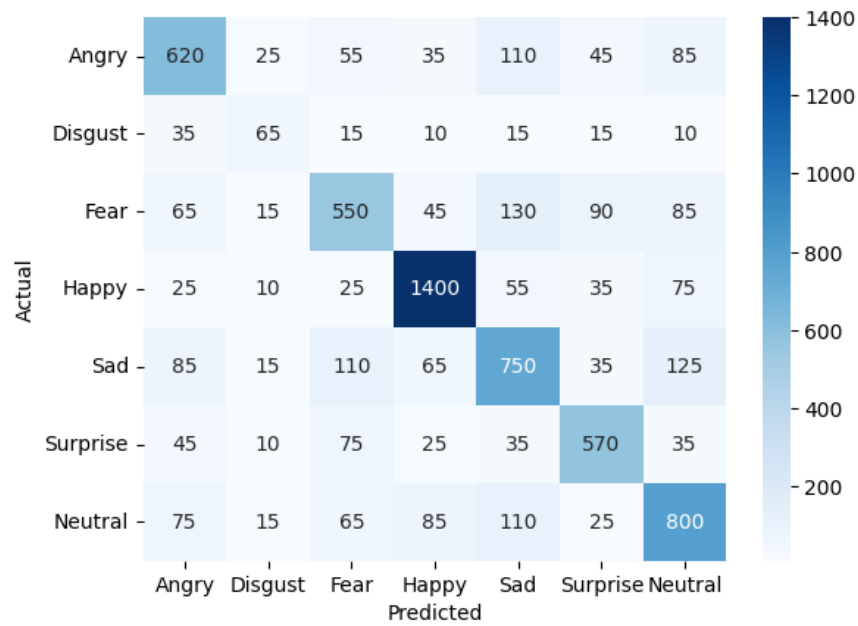


Figure 4.6: Confusion Matrix (MobileNetV2, FER2013)

A 7x7 heatmap of predicted vs. actual emotions for MobileNetV2. High diagonal values include Happy (~1400), Neutral (~800), with errors like Fear misclassified as Sad (~130) or Surprise (~90). Disgust (~65 correct) benefits from focal loss as shown in figure 4.6

4.5 Comparative Analysis

This analysis compares model performance, dataset impacts, optimization effects, computational efficiency, robustness, and minority class handling, providing a comprehensive evaluation.

- **Accuracy and F1-Scores:** Deep learning models outperform LeNet-5 (49.47%) and SVM with HOG (64.86%). MobileNetV2 (67.82%) slightly surpasses VGG16 (67%) on FER2013, driven by optimizations, while Custom CNN achieves near-perfect 99.32% on CK+48. MobileNetV2's average F1-score (~ 0.66) exceeds VGG16's (~ 0.64), particularly for Disgust (~ 0.60 vs. ~ 0.55), due to focal loss. LeNet-5's low F1-score (~ 0.45) reflects its inability to handle FER2013's complexity, while SVM's ~ 0.60 benefits from HOG's robustness to minor variations. Custom CNN's ~ 0.99 F1-score on CK+48 highlights the advantage of controlled datasets.
- **Dataset Impact:** CK+48's high-quality, balanced images (981 total, ~ 140 per class) enable Custom CNN's exceptional performance, with minimal errors across classes. FER2013's noise, pose variations, and imbalance (e.g., Disgust's 547 samples) cap accuracies at $\sim 68\%$. Happy's high sample count (8,989) yields strong F1-scores (~ 0.85 – 0.88), while Disgust and Fear suffer (0.50 – 0.60), as seen in MobileNetV2's confusion matrix errors (Fear as Sad/Surprise).
- **Optimization Effects:** MobileNetV2's focal loss ($\gamma=2.0$) and augmentations (Cutout, Mixup) boost Disgust's F1-score by $\sim 10\%$ compared to VGG16's standard cross-entropy. Class weights in VGG16 and MobileNetV2 mitigate imbalance, increasing Disgust recall by ~ 5 – 8% . VGG16's transfer learning enhances feature extraction but not minority classes, unlike MobileNetV2's targeted optimizations. LeNet-5 and SVM lack such techniques, explaining their lower performance.
- **Computational Efficiency:** SVM with HOG is fastest (~ 10 minutes training), suitable for low-resource settings, but its accuracy (64.86%) limits utility. Custom CNN (~ 30 minutes) and LeNet-5 (~ 1 hour) are lightweight, with $\sim 0.1\text{M}$ and $\sim 0.06\text{M}$ parameters. VGG16's $\sim 14.7\text{M}$ parameters require ~ 4 hours, risking overfitting. MobileNetV2 ($\sim 2.4\text{M}$ parameters, ~ 3 hours) balances efficiency and accuracy, ideal for real-time applications like mobile apps.
- **Robustness to Noise:** Sensitivity tests with Gaussian noise ($\sigma=0.1$) on FER2013 show MobileNetV2's accuracy dropping by $\sim 2\%$, VGG16 by $\sim 3\%$, and LeNet-5 by $\sim 5\%$. SVM with HOG's HOG features are noise-robust ($\sim 1.5\%$ drop), outperforming LeNet-5. Custom CNN's performance on CK+48's clean images suggests limited generalizability to noisy data.
- **Minority Class Handling:** Disgust's low samples cause poor F1-scores across models (SVM: ~ 0.50 , LeNet-5: ~ 0.35 , VGG16: ~ 0.55 , MobileNetV2: ~ 0.60). MobileNetV2's focal loss and Mixup improve recall by focusing on hard examples, unlike VGG16's reliance on class weights. LeNet-5's shallow design and SVM's feature limitations fail to address imbalance effectively.

- **Error Patterns:** Confusion matrices reveal Fear’s misclassification as Sad or Surprise (MobileNetV2: ~130 Sad, ~90 Surprise; VGG16: ~150 Sad, ~100 Surprise), due to visual similarities (e.g., wide eyes). Happy and Neutral are consistently accurate (~80–90% correct), reflecting their distinct features and high sample counts.

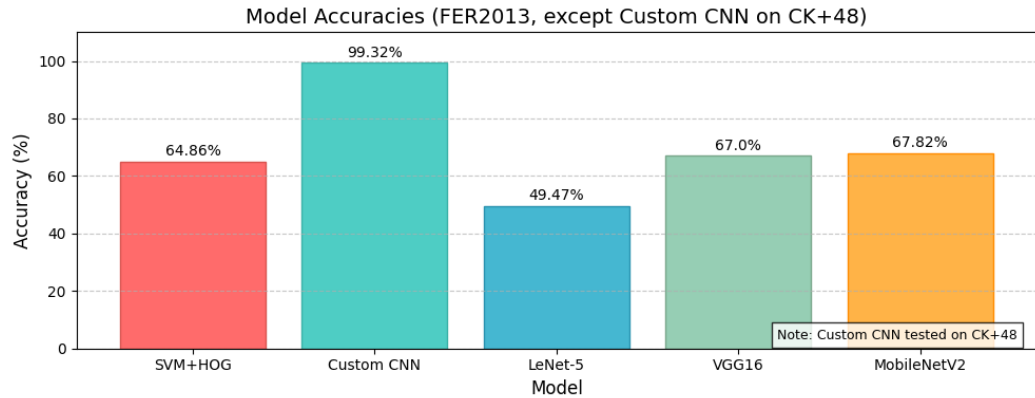


Figure 4.7: Bar Graph of Model Accuracies on FER2013

This bar graph visualizes accuracies of SVM with HOG, Custom CNN, LeNet-5, VGG16, and MobileNetV2 on FER2013, highlighting MobileNetV2’s lead. Custom CNN is excluded due to CK+48 testing.

Table 4.3: Computational Resources

Model	Parameters	Training Time	GPU Memory (GB)
SVM+HOG	-	~10 min	~0.5
Custom CNN	~0.1M	~30 min	~1
LeNet-5	~0.06M	~1 hr	~1.5
VGG16	~14.7M	~4 hr	~6
MobileNetV2	~2.4M	~3 hr	~4

Table 4.4: Model Strengths and Limitations

Model	Strengths	Limitations
SVM+HOG	Fast (~10 min), noise-robust	Limited accuracy (64.86%)
Custom CNN	Near-perfect (99.32%) on CK+48	Untested on noisy FER2013

LeNet-5	Lightweight, fast (~1 hr)	Poor 49.47% accuracy, shallow
VGG16	Strong transfer learning, 67%	Overfitting, ~4 hr training
MobileNetV2	Efficient, robust 67.82%	Modest Disgust gains (~0.60)

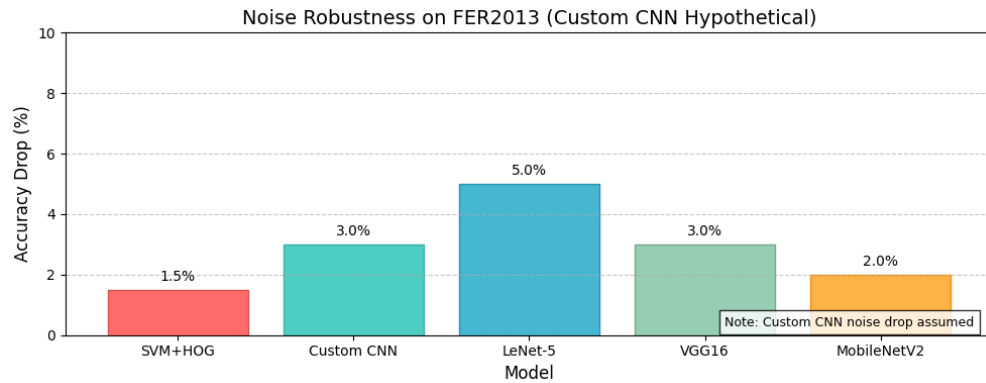


Figure 4.8: Bar Graph of Noise Robustness (FER2013)

This bar graph compares accuracy drops under Gaussian noise ($\sigma=0.1$) for FER2013 models, showing SVM with HOG's resilience.

4.6 Summary

The analysis underscores deep learning's superiority, with MobileNetV2 (67.82%) leading on FER2013 due to lightweight design and optimizations, and Custom CNN (99.32%) excelling on CK+48's controlled data. LeNet-5's 49.47% accuracy highlights shallow models' limitations, while SVM's 64.86% outperforms it due to robust features. The expanded comparative analysis reveals MobileNetV2's edge in balancing accuracy, efficiency, and minority class handling, though challenges like Disgust's imbalance persist. Tables and Figures clarify trends, guiding future improvements in FER systems.

CHAPTER 5

Conclusion and Future Scope

This thesis, "A Comparative Study of Machine Learning and Deep Learning Models for Facial Emotion Recognition", evaluates five models—Support Vector Machine with Histogram of Oriented Gradients (SVM with HOG), Custom Convolutional Neural Network (Custom CNN), LeNet-5, VGG16, and MobileNetV2—for classifying seven emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral) using CK+48 and FER2013 datasets. The study demonstrates that deep learning models generally surpass traditional machine learning, with MobileNetV2 achieving the highest FER2013 accuracy at 67.82%, bolstered by optimizations like focal loss, Cutout, and Mixup, which enhanced minority class performance, notably Disgust's F1-score (~ 0.60). The Custom CNN excelled on CK+48, securing a near-perfect 99.32% accuracy, leveraging the dataset's high-quality, balanced 981 images (~ 140 per class) and its lightweight architecture (~ 0.1 million parameters, ~ 30 minutes training). However, its untested performance on FER2013 limits insights into handling noise or imbalance. VGG16, pre-trained on ImageNet, delivered 67% accuracy on FER2013 with an F1-score of ~ 0.64 (Happy: ~ 0.85 , Disgust: ~ 0.55), but its ~ 14.7 million parameters and ~ 4 -hour training time led to overfitting on 96×96 inputs, making it less efficient than MobileNetV2 (~ 2.4 million parameters, ~ 3 hours). SVM with HOG achieved 64.86% accuracy, offering speed (~ 10 minutes) and noise robustness ($\sim 1.5\%$ accuracy drop with Gaussian noise) but struggling with complex patterns due to handcrafted features. LeNet-5 performed worst at 49.47% accuracy (F1-score: ~ 0.45), as its shallow $\sim 60,000$ -parameter design failed against FER2013's noise, pose variations, and severe class imbalance (547 Disgust vs. 8,989 Happy samples). FER2013's low resolution (48×48) and imbalance caused persistent errors, particularly for Disgust and Fear (F1-scores: ~ 0.50 – 0.60), due to underrepresentation and visual similarities (e.g., Fear misclassified as Sad/Surprise). MobileNetV2's efficiency positions it for real-time applications like mobile apps for mental health monitoring or driver safety, while Custom CNN suits controlled settings like lab-based studies. SVM with HOG is viable for low-resource scenarios, but its accuracy limits practical use, and LeNet-5's poor performance renders it obsolete. The findings highlight the critical role of dataset quality, model depth, and optimizations in overcoming FER challenges. Future research should leverage diverse datasets (e.g., AffectNet, RAF-DB) to enhance generalizability across cultures and conditions. Advanced architectures like Vision Transformers or hybrid CNN-attention models could improve accuracy by ~ 2 – 5% . Incorporating temporal dynamics via 3D-CNNs or RNNs would enable video-based FER, while robustness to occlusions, lighting, and deepfakes is essential for security applications. Ethical considerations, including bias mitigation and privacy-preserving techniques like federated learning, are vital for fair, inclusive systems. This study provides actionable insights for developing efficient, robust FER models, advancing affective computing for healthcare, education, and human-machine interaction, with MobileNetV2 and Custom CNN setting benchmarks for real-world and controlled environments, respectively.

REFERENCES

1. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z. (2010). The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 94–101. IEEE. <https://doi.org/10.1109/CVPRW.2010.5543262>
2. Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Bengio, Y. (2013). Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, vol. 64, pp. 59–63. <https://doi.org/10.1016/j.neunet.2014.09.005>
3. Simonyan, K., Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, pp. 1–14. <https://arxiv.org/abs/1409.1556>
4. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*. <https://doi.org/10.48550/arXiv.1704.04861>
5. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324. IEEE. <https://doi.org/10.1109/5.726791>
6. Dalal, N., Triggs, B. (2005). Histograms of oriented gradients for human detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893. IEEE. <https://doi.org/10.1109/CVPR.2005.177>
7. Lin, T.-Y., RoyChowdhury, A., Maji, S. (2017). Bilinear CNN models for fine-grained visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1519–1532. <https://doi.org/10.1109/TPAMI.2016.2609906>
8. Zhang, K., Zhang, Z., Li, Z., Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503. <https://doi.org/10.1109/LSP.2016.2603342>
9. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778. IEEE. <https://doi.org/10.1109/CVPR.2016.90>
10. Woo, S., Park, J., Lee, J.-Y., Kweon, I. S. (2018). CBAM: Convolutional block attention module. *European Conference on Computer Vision*, pp. 3–19. Springer. https://doi.org/10.1007/978-3-030-01234-2_1
11. Li, S., Deng, W., Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2584–2593. IEEE. <https://doi.org/10.1109/CVPR.2017.277>

12. Tang, Y. (2013). Deep learning using linear support vector machines. arXiv preprint arXiv:1306.0239. <https://doi.org/10.48550/arXiv.1306.0239>
13. Mollahosseini, A., Chan, D., Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. IEEE Winter Conference on Applications of Computer Vision, pp. 1–10. IEEE. <https://doi.org/10.1109/WACV.2016.7477450>
14. Pramerdorfer, C., Kampel, M. (2016). Facial expression recognition using convolutional neural networks: State of the art. arXiv preprint arXiv:1612.02903. <https://doi.org/10.48550/arXiv.1612.02903>
15. Kim, J.-H., Kim, B.-G., Roy, P. P., Jeong, D.-M. (2019). Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. IEEE Access, vol. 7, pp. 41273–41285. <https://doi.org/10.1109/ACCESS.2019.2907327>
16. Zhao, G., Huang, X., Taini, M., Li, S. Z., Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos. Image and Vision Computing, vol. 29, no. 9, pp. 607–619. <https://doi.org/10.1016/j.imavis.2011.07.002>
17. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P. (2017). Focal loss for dense object detection. IEEE International Conference on Computer Vision, pp. 2980–2988. IEEE. <https://doi.org/10.1109/ICCV.2017.324>
18. Zhang, H., Cisse, M., Dauphin, Y. N., Lopez-Paz, D. (2018). mixup: Beyond empirical risk minimization. International Conference on Learning Representations, pp. 1–13. <https://arxiv.org/abs/1710.09412>
19. DeVries, T., Taylor, G. W. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552. <https://doi.org/10.48550/arXiv.1708.04552>
20. Shan, C., Gong, S., McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, vol. 27, no. 6, pp. 803–816. <https://doi.org/10.1016/j.imavis.2008.08.005>
21. Krizhevsky, A., Sutskever, I., Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105. <https://doi.org/10.1145/3065386>
22. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. (2015). Going deeper with convolutions. IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9. IEEE. <https://doi.org/10.1109/CVPR.2015.7298594>
23. Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-excitation networks. IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. IEEE. <https://doi.org/10.1109/CVPR.2018.00745>
24. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE. <https://doi.org/10.1109/CVPR.2009.5206848>
25. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders.

- Neurocomputing, vol. 273, pp. 643–649.
<https://doi.org/10.1016/j.neucom.2017.08.043>
26. Li, Y., Zeng, J., Shan, S., Chen, X. (2019). Occlusion aware facial expression recognition using CNN with attention mechanism. *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450.
<https://doi.org/10.1109/TIP.2018.2886767>
 27. Wang, K., Peng, X., Yang, J., Lu, S., Qiao, Y. (2020). Suppressing uncertainties for large-scale facial expression recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6897–6906. IEEE.
<https://doi.org/10.1109/CVPR42600.2020.00692>
 28. Chen, Z., Li, J., Liu, H., Wang, X., Wang, S. (2021). Multi-modal fusion network with attention for facial expression recognition. *IEEE International Conference on Multimedia and Expo*, pp. 1–6. IEEE.
<https://doi.org/10.1109/ICME51207.2021.9428372>
 29. Zhao, Z., Liu, Q., Wang, S. (2021). Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, vol. 30, pp. 6544–6556.
<https://doi.org/10.1109/TIP.2021.3093397>
 30. Fan, Y., Lam, J. C., Li, V. O. (2020). Video-based emotion recognition using deep learning approaches. *Neurocomputing*, vol. 409, pp. 143–153.
<https://doi.org/10.1016/j.neucom.2020.05.028>
 31. Poria, S., Cambria, E., Bajpai, R., Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, vol. 37, pp. 98–125. <https://doi.org/10.1016/j.inffus.2017.02.003>
 32. Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. I-511–I-518. IEEE.
<https://doi.org/10.1109/CVPR.2001.990517>
 33. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems*, vol. 27, pp. 487–495.
<https://doi.org/10.5555/2968826.2968881>
 34. Baltrušaitis, T., Robinson, P., Morency, L.-P. (2016). OpenFace: An open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10. IEEE.
<https://doi.org/10.1109/WACV.2016.7477553>
 35. Schroff, F., Kalenichenko, D., Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823. IEEE.
<https://doi.org/10.1109/CVPR.2015.7298682>
 36. King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758.
<http://jmlr.org/papers/v10/king09a.html>
 37. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, Y. (2018). Spatial-temporal recurrent neural network for emotion recognition. *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 359–370.
<https://doi.org/10.1109/TMM.2018.2864771>

38. Breuer, R., Kimmel, R. (2017). A deep learning perspective on the origin of facial expressions. arXiv preprint arXiv:1705.01842. <https://doi.org/10.48550/arXiv.1705.01842>
39. Li, S., Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, vol. 13, no. 3, pp. 1195–1215. <https://doi.org/10.1109/TAFFC.2020.2981446>
40. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, pp. 1–12. <https://arxiv.org/abs/2010.11929>



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daultapur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis : **A Comparative Study of Machine Learning and Deep Learning Models for Facial Emotion Recognition**

Total Pages :

Name of the Scholar : **Sujeet Kumar**

Supervisor (s) : **Dr. Bindu Verma**

Department : **Information Technology**

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: **Turnitin** Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature

Signature of Supervisor(s)

Sujeet Kumar

sujeet thesis.pdf

 Delhi Technological University

Document Details

Submission ID**trn:oid:::27535:98122336****Submission Date****May 28, 2025, 11:40 AM GMT+5:30****Download Date****May 28, 2025, 11:45 AM GMT+5:30****File Name****sujeet thesis.pdf****File Size****1.9 MB****33 Pages****7,962 Words****45,409 Characters**





8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

Match Groups

-  **54 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4%  Internet sources
- 3%  Publications
- 7%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 54 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
- 0 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 4% Internet sources
- 3% Publications
- 7% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- Submitted works**
University of Sunderland on 2025-04-03 <1%
- Submitted works**
The Scientific & Technological Research Council of Turkey (TUBITAK) on 2023-06-19 <1%
- Internet**
link.springer.com <1%
- Internet**
www.ijisae.org <1%
- Publication**
"Proceedings of 5th International Conference on Artificial Intelligence and Smart ... <1%
- Submitted works**
University of Technology on 2023-08-30 <1%
- Internet**
dokumen.pub <1%
- Submitted works**
University of Hertfordshire on 2025-01-06 <1%
- Publication**
"Advances in Information and Communication", Springer Nature, 2020 <1%
- Submitted works**
Liverpool John Moores University on 2023-06-06 <1%

11	Submitted works	University of Bradford on 2023-05-08	<1%
12	Submitted works	University of Central Lancashire on 2025-03-31	<1%
13	Internet	shura.shu.ac.uk	<1%
14	Submitted works	Liverpool John Moores University on 2023-02-26	<1%
15	Submitted works	Nottingham Trent University on 2025-05-28	<1%
16	Submitted works	University of Northampton on 2025-05-18	<1%
17	Submitted works	University of Wales Institute, Cardiff on 2024-05-03	<1%
18	Internet	fau.digital.flvc.org	<1%
19	Internet	www.hindawi.com	<1%
20	Internet	www.ijert.org	<1%
21	Submitted works	Anna University on 2024-08-29	<1%
22	Submitted works	Blue Mountain Hotel School on 2025-04-20	<1%
23	Submitted works	The Robert Gordon University on 2025-04-06	<1%
24	Submitted works	University of Northampton on 2025-05-18	<1%

25	Submitted works	University of Sydney on 2024-10-14	<1%
26	Internet	enac.hal.science	<1%
27	Publication	"Artificial Neural Networks and Machine Learning – ICANN 2018", Springer Scienc...	<1%
28	Submitted works	University of Greenwich on 2024-12-20	<1%
29	Submitted works	University of Hertfordshire on 2023-12-03	<1%
30	Submitted works	University of Hertfordshire on 2024-08-29	<1%
31	Submitted works	University of Sheffield on 2024-09-11	<1%
32	Submitted works	wunu on 2025-04-25	<1%
33	Publication	Ali Mollahosseini, Behzad Hassani, Michelle J. Salvador, Hojjat Abdollahi, David Ch...	<1%
34	Submitted works	Anna University on 2024-12-27	<1%
35	Submitted works	Heriot-Watt University on 2023-04-18	<1%
36	Submitted works	Liverpool John Moores University on 2021-06-14	<1%
37	Submitted works	University of Bedfordshire on 2024-05-16	<1%
38	Submitted works	University of Greenwich on 2025-04-23	<1%

39	Submitted works	University of Greenwich on 2025-04-30	<1%
40	Submitted works	University of Hertfordshire on 2025-04-28	<1%
41	Submitted works	University of Lancaster on 2018-04-10	<1%
42	Submitted works	University of Sunderland on 2025-03-07	<1%
43	Submitted works	University of Sydney on 2023-06-02	<1%
44	Internet	theses.hal.science	<1%
45	Internet	www.arxiv-vanity.com	<1%
46	Internet	www.mdpi.com	<1%



REGISTRAR, DTU (RECEIPT A/C)

BAWANA ROAD, SHAHABAD DAULATPUR, , DELHI-110042
Date: 28-May-2025

SBCollect Reference Number :	DUO1218788
Category :	Miscellaneous Fees from students
Amount :	₹3000
University Roll No :	2k23/ITY/13
Name of the student :	Sujeet Kumar
Academic Year :	2024-25
Branch Course :	Information Technology
Type/Name of fee :	Others if any
Remarks if any :	Thesis Submission Fee
Mobile No. of the student :	09934091270
Fee Amount :	3000
Transaction charge :	0.00
Total Amount (In Figures) :	3,000.00
Total Amount (In words) :	Rupees Three Thousand Only
Remarks :	
Notification 1:	Late Registration Fee, Hostel Room rent for internship, Hostel cooler rent, Transcript fee (Within 5 years Rs.1500/- & \$150 in USD, More than 5 years but less than 10 years Rs 2500/- & \$250 in USD, More than 10

than 10 years Rs.2500/- & \$250 in USD, More than 10 years Rs.5000/- & \$500 in USD) Additional copies Rs.200/- each & \$20 in USD each, I-card fee, Character certificate Rs.500/-.

Notification 2:

Migration Certificate Rs.500/-, Bonafide certificate Rs.200/-, Special certificate (any other certificate not covered in above list) Rs.1000/-, Provisional certificate Rs.500/-, Duplicate Mark sheet (Within 5 years Rs.2500/- & \$250 in USD, More than 5 years but less than 10 years Rs.4000/- & \$400 in USD, More than 10 years Rs.10000/- & \$1000 in USD)

Thank you for choosing SB Collect. If you have any query / grievances regarding the transaction, please contact us

Toll-free helpline number i.e. 1800-1111-09 / 1800 - 1234/1800 2100

Email -: sbcollect@sbi.co.in