Kaunas University of Technology

Faculty of Informatics

Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low Data Regimes

Master's Final Degree Project

Ananthakrishnan Thuruthel Murali

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low Data Regimes

Master's Final Degree Project

Artificial Intelligence in Computer Science (6211BX007)

Ananthakrishnan Thuruthel Murali

Project author

Prof. Dr. Armantas Ostreika

Supervisor

Assoc. Prof. Marius Pivoras

Reviewer

Kaunas, 2026

Kaunas University of Technology

Faculty of Informatics

Ananthakrishnan Thuruthel Murali

Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low Data Regimes

Declaration of Academic Integrity

I confirm the following:

1. I have prepared the final degree project independently and honestly without any violations of the copyrights or other rights of others, following the provisions of the Law on Copyrights and Related Rights of the Republic of Lithuania, the Regulations on the Management and Transfer of Intellectual Property of Kaunas University of Technology (hereinafter – University) and the ethical requirements stipulated by the Code of Academic Ethics of the University;

2. All the data and research results provided in the final degree project are correct and obtained legally; none of the parts of this project are plagiarised from any printed or electronic sources; all the quotations and references provided in the text of the final degree project are indicated in the list of references;

3. I have not paid anyone any monetary funds for the final degree project or the parts thereof unless required by the law;

4. I understand that in the case of any discovery of the fact of dishonesty or violation of any rights of others, the academic penalties will be imposed on me under the procedure applied at the University; I will be expelled from the University and my final degree project can be submitted to the Office of the Ombudsperson for Academic Ethics and Procedures in the examination of a possible violation of academic ethics.

Ananthakrishnan Thuruthel Murali

Confirmed electronically

Kaunas University of Technology

Artificial Intelligence in Computer Science

Task of the Master's final degree project

Topic of the project	Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low data Regimes
Requirements and conditions (title can be clarified, if needed)	
Supervisor	(position, name, surname, signature of the supervisor) (date)

Ananthakrishnan Thuruthel Murali. Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low data Regimes. Master's Final Degree Project / supervisor prof. dr. Armantas Ostreika; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Computer Science, Informatics (B01).

Keywords: facial expression recognition, transfer learning, MobileNetV2, subnetwork selection, small dataset.

Kaunas, 2026. 55 p.

Summary

Facial expression recognition is becoming an important topic with the introduction of self-driving cars, or whether it be to find the mental state of a patient in a hospital or to find out if a person is lying. A key challenge in this field is achieving reliable performance under small dataset conditions, where standard deep learning methods are prone to overfitting.

This study investigated the use of MobileNetV2 based subnetwork selection as a parameter efficient approach to FER under limited data conditions. A global search strategy was developed to identify the optimal subnetwork, defined as blocks 1 through k, where $k \in \{3 \ldots . .17\}$, by evaluating the candidate paths using a lightweight logistic regression probe on a subject independent subset of the data. The selected subnetwork was then finetuned and evaluated against a full MobileNetV2 baseline across FER2013 and CK+ datasets. The exhaustive search was able to identify blocks 1-13 as the optimal subnetwork, achieving a 55.8% parameter reduction (2.23M to 0.99M). On FER2013, the subnetwork produced no statically significant accuracy at 10-20% data fractions while reducing the train-validation accuracy gap by 35-48% across all evaluated data fractions across all evaluated data fractions, which confirms a significant lower overfitting than the full model. On the CK+ dataset under subject independent five-fold cross validation, the subnetwork achieved a 90.47% accuracy compared to 88.58% for the baseline, with macro F1 improving from 0.834 to 0.862. These results indicate that a carefully selected compact subnetwork can produce competitive accuracy in a moderate low data setting while giving significant advantage in parameter efficiency and overfitting resistance.

Use of Artificial Intelligence Tools: Generative Ai tools were used during preparation of this thesis for grammar checking, code suggestions and language correction. All the research, analysis and experiments done are author's own work.

Ananthakrishnan Thuruthel Murali. Parametrais efektyvus „MobileNetV2“ potinklio parinkimas veido išraiškų atpažinimui mažo duomenų srauto režimuose. Magistro baigiamasis projektas / vadovas prof. dr. Armantas Ostreika; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos mokslai, Informatika (B01).

Reikšminiai žodžiai: veido išraiškos atpažinimas, mokymosi perkėlimas, MobileNetV2, potinklio pasirinkimas, mažas duomenų rinkinys.

Kaunas, 2026. 55 p.

Santrauka

Veido išraiškų atpažinimas tampa svarbia tema atsiradus savaeigiams automobiliams, tiek norint nustatyti paciento psichinę būseną ligoninėje, tiek norint išsiaiškinti, ar žmogus meluoja. Pagrindinis šios srities iššūkis yra užtikrinti patikimą veikimą esant mažiems duomenų rinkiniams, kai standartiniai gilaus mokymosi metodai yra linkę į perteklinį pritaikymą.

Šiame tyrime buvo nagrinėjamas „MobileNetV2“ pagrįsto potinklio parinkimo naudojimas kaip parametrų požiūriu efektyvus FER metodas esant ribotiems duomenų kiekiams. Buvo sukurta globali paieškos strategija, skirta nustatyti optimalų potinklį, apibrėžtą kaip blokai nuo 1 iki k, kur k $\in\{3\ldots..17\}$, įvertinant kandidatų kelius naudojant lengvą logistinės regresijos zondą su nuo subjekto nepriklausomu duomenų pogrupiu. Pasirinktas potinklis buvo tiksliai suderintas ir įvertintas pagal visą „MobileNetV2“ bazinę liniją FER2013 ir CK+ duomenų rinkiniuose. Išsami paieška leido nustatyti blokus nuo 1 iki 13 kaip optimalų potinklį, pasiekus 55,8 % parametrų sumažinimą (nuo 2,23 mln. iki 0,99 mln.). FER2013 modelyje potinklis nesukūrė statistiškai reikšmingo tikslumo esant 10–20 % duomenų dalims, tuo tarpu traukinio patvirtinimo tikslumo skirtumas sumažėjo 35–48 % visose įvertintose duomenų dalyse, o tai patvirtina žymiai mažesnį perteklinį pritaikymą nei visas modelis. CK+ duomenų rinkinyje, atliekant nepriklausomą penkių kartų kryžminį patvirtinimą, potinklis pasiekė 90,47 % tikslumą, palyginti su 88,58 % baziniame lygmenyje, o makro F1 pagerėjo nuo 0,834 iki 0,862. Šie rezultatai rodo, kad kruopščiai parinktas kompaktiškas potinklis gali užtikrinti konkurencingą tikslumą esant vidutiniškai mažam duomenų kiekiui, tuo pačiu suteikdamas didelį pranašumą parametrų efektyvumo ir atsparumo pertekliniam pritaikymui srityse.

Dirbtinio intelekto įrankių naudojimas: Rengiant šį darbą gramatikos tikrinimui, kodo pasiūlymams ir kalbos taisymui buvo naudojami generatyvinio dirbtinio intelekto įrankiai. Visi atlikti tyrimai, analizė ir eksperimentai yra autoriaus darbas.

Table of contents

List of figures

List of Tables

List of abbreviations and terms

Abbreviations:

FER – Facial Expression Recognition.

ReLU – Rectified Linear Unit

DCNN - Deep Convolutional Neural Network

Train-val – Train - Validation

Introduction

Project novelty and relevance

FER has become an important field within computer vision, by enabling different applications such as emotion detection, virtual reality interaction, user experience optimization and security systems. An accurate FER system can greatly improve human-computer interactions which will make the technology more adaptive and empathetic. Despite the increase in number of deep learning techniques, which has significantly advanced the field, many existing approaches rely a lot on large-scale annotated datasets to achieve high performance. This presents a huge barrier in situations where acquiring such datasets is impractical due to constraints like cost, privacy concerns or domain-specific requirements.

In this situation, the challenge of performing FER with small dataset is both critical and underexplored. Small datasets often cause issues such as overfitting, reduced generalization and poor performance in real-world environments. These challenges have to be addressed to expand the accessibility and applicability of FER systems, like personalized healthcare or industrial application.

This research is mainly focused on evaluating and improving an FER system for situations where the data will be limited. By leveraging methods like transfer learning, data augmentation and lightweight convolutional neural networks, the study aims to bridge the gap between the high-performance FER systems and practical deployment in situations where the data is limited. The proposed method not only reduces the challenges caused by small datasets but also contributes to a broader field by offering scalability and efficient solutions.

Problem Statement

Facial expression recognition can be formulated as a multi-class image classification problem. For a given dataset

$$D = \{(x_i, y_i)\}_{i=1}^{N}, \quad (1)$$

Where x_i, represents a facial image and $y_i \in \{1,..,7\}$ denotes the corresponding class. The objective is to make a classifier f_θ learn, so that it can maximize the classification performance on unseen data.

In practical scenarios, the number of labelled samples N, is limited. This leads to overfitting and poor generalization. This thesis addresses the problem of improving the performance of facial expression recognition under the constraint of small dataset, by leveraging transfer learning and optimization of the architecture.

Aim and objectives

The main aim is to evaluate and improve an FER model that can accurately detect emotion using small dataset. To achieve this goal, the following objectives were set:

1. Find the existing solutions through literature analysis.
2. Design a new architecture which can improve FER model.
3. Develop the model or fine tune an existing model
4. Evaluate the results with existing models

Research Objective

The research objective is parameter efficient subnetwork selection within a pretrained MobileNetV2 backbone for facial expression recognition under limited data conditions.

Document Structure

The document outlines the development and evaluation of and FER system with small dataset. The thesis paper extends over three chapters beginning with literature analysis of previous works. The first chapter will explore the challenges in FER by focusing on addressing the limitations of small datasets through techniques such as data augmentation, transfer learning and lightweight neural networks. The second chapter details the proposed methods and architecture design and the third chapter presents the experimental results and discussion. The thesis then concludes with separate section for final observations limitations and conclusions.

1. Analysis of Facial Expression Recognition Using Small Dataset

This chapter will focus on the different methods and algorithms used for FER with small dataset and the different techniques used for the facial expression detection. It will also discuss the different preprocessing and data augmentation methods. The study will also analyse Transfer learning and other methods.

1.1. Existing Solutions for FER With Small Dataset

A few techniques have been applied for FER using small dataset. One such method is by applying transfer learning [1], where they used AlexNet and VGG-CNN-M-2048 models pretrained on ImageNet dataset to perform a two-stage fine-tuning. The first stage fine tuning was done on a 2013 facial expression dataset, containing 28k/32k low resolution images which were collected from the internet. In the second stage the fine tuning was done based on the training part of EmotiW dataset which consist of static facial expression recognition extracted from movies [1].

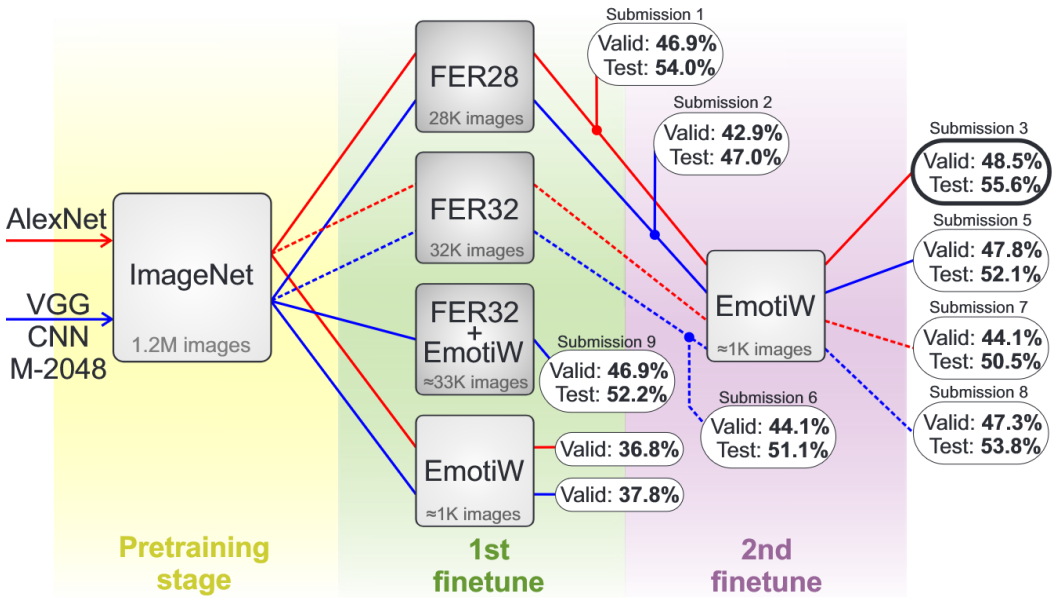

Figure 1. Shows the different stages [1]

While the second stage fine tuning only had a marginal impact the first stage fine tuning helped significantly to achieving better result which is using the FER-2013 dataset. Another researcher used Transfer learning [2] in a single stage using the MobileNetV2 model which is a 53-Layered DCNN, which is also pretrained on images from ImageNet database. Fine tuning was done on JAFFE dataset and resulted in an accuracy of 85.54%. The work proposed in this paper [3] proposes an advanced version of MobileNetV2 for FER, that incorporates the attention mechanism along with the inverted residual blocks, achieving an accuracy of 70.21% on FER 2013 and 98012% on CK+. While this shows that the improved architecture of the lightweight model can yield strong performance, the approach increases model complexity through the addition of attention modules rather than reducing it.

This paper [4] on the other hand combined active learning with contrastive self-supervised pretraining. The combining was done to address the phenomenon of “cold start”. The cold start happens when the initial labelled data set is too small. Their method was able to achieve an accuracy of 67% on FER13. When looking through the results of the proposed methods from papers [1, 2, 4], it can be observed that the method from paper [2]has a higher accuracy as when compared to the rest of the papers.

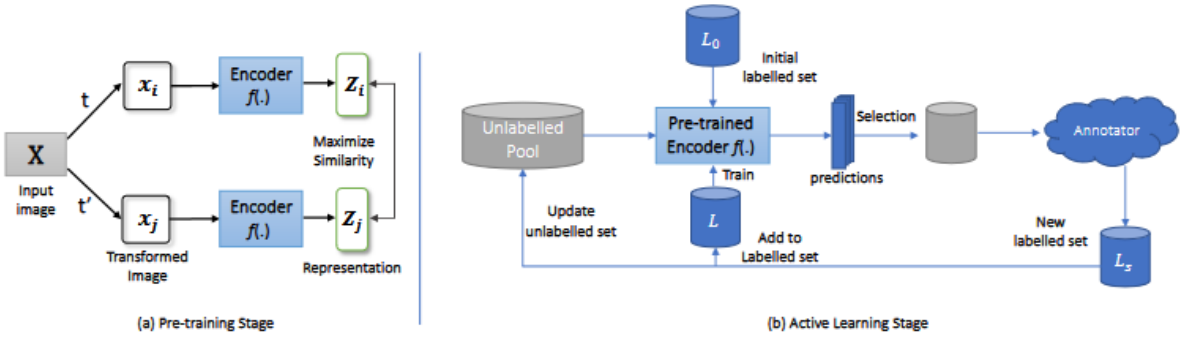

Figure 2. Gives overview of the pre-training and active learning stage [4]

1.2. Existing Solutions for FER

Existing methods for facial expression detection predominantly focus on deep learning methods to extract features and using classification pipelines. The main fundamental design decision for these approaches involve, defining a fixed set of target emotion categories, including anger, fear .etc. These categories will guide both the network and architecture design and loss calculation, as the model has to learn discriminative representations that can separate visually subtle facial cues.

1.2.1. Existing Methods

In order to detect the different emotions the first step would be to identify which of the main expression have to be classified so that the model can identify those expressions based off from the image and for this purpose, it is first classified into seven categories like happy, sad, fearful, anger, surprised disgust and neutral [5, 6]. To detect the expressions, multiple models are used. Some of the popular models are Resnet, which is known for its ability to handle complex visual tasks. It uses a series of convolutional layers in residual blocks to learn and extract detailed features from the input data while also reducing the gradient vanishing problem. They extracted the features using CNN, BN and activation function ReLU.[6]The accuracy of the experiment was around 95%. A similar experiment was also conducted with Resnet-50 model, but it was done using transfer learning [5],where the model was fine-tuned which gave an accuracy of 99%.

Figure 3. Example for seven expressions[6]

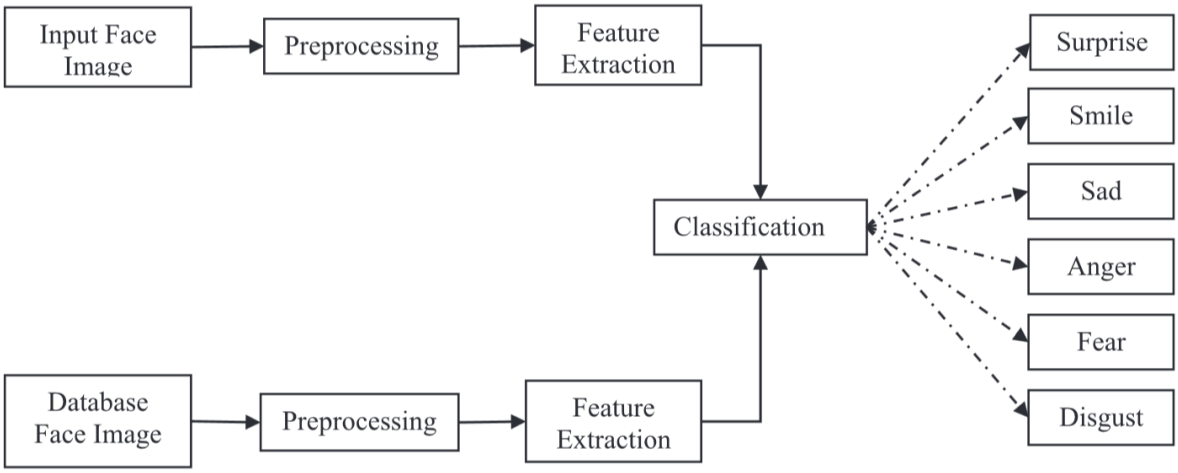

Figure 4. Diagram of an expression detection system [7]

While the method mentioned above was using ResNet architecture another paper proposed the use of AlexNet, the paper proposed a joint optimization method using softmax loss and improved island loss into the AlexNet-Emotion network. The softmax helps the model to learn the expressions while the improved island loss function can make the expression features more descriptive and compact [8].

1.3. Unsupervised Learning

This subsection briefly goes through the different unsupervised learning methods used for FER process. Unsupervised learning is where the model aims to understand the data's structure by grouping into similar data points or to transform it into new representation. [9]

1.3.1. Transfer Learning

While the methods for facial expression recognition with small dataset is less, the most popular or common once that have been used is transfer learning. So, what exactly is transfer learning, according to [10]the paper transfer learning aims to make the model work better on a new task by using the knowledge already learned from a similar related task Figure gives some intuitive examples about transfer learning. So instead of starting from scratch and requiring a lot of data transfer learning will just reuse parts of a model that was trained on a large dataset. Some of the applications was already discussed in the first section of existing solutions.

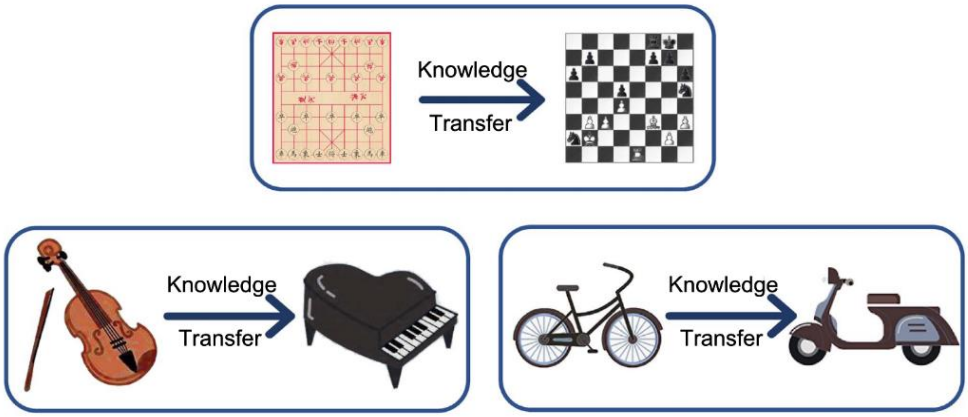

Figure 5. Intuitive examples about transfer learning[10]

1.3.2. Meta Learning

Meta Learning has seen significant advancement and increased application in recent years. It focuses on learning to learn concept, that is using experiences across multiple tasks to improve future learning process[11]. A research team used Meta Transfer Learning through PathNet, a specialized network architecture that utilizes the pathways for efficient transfer learning. PathNet identifies the most optimal pathways and fixes them to prevent forgetting and reinitializes non-pathway parameters. This method was applied to audio datasets, and it outperformed the traditional finetuning techniques and had an overall accuracy of 94% overall [12].

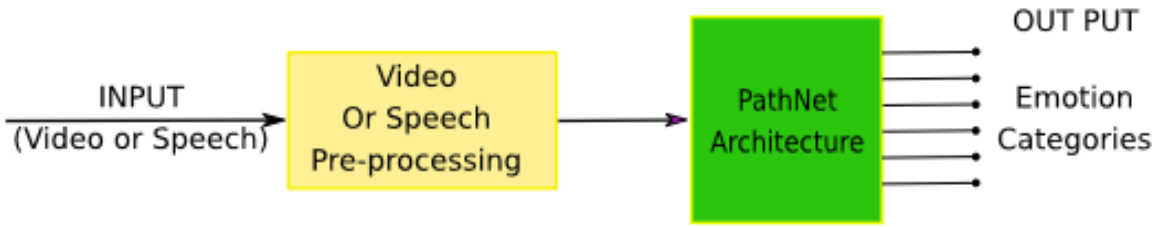

Figure 6. Pipeline of the method used [12]

NAS “Neural Architecture Search”, is another method which represents a broader class of method that is used to automatically identify the most efficient network structure for a given task. [13] applied NAS using genetic algorithm specifically for FE, demonstrating that the automated architecture search can identify competitive architectures. However, NAS methods typically require training and evaluating hundreds of candidate architecture, making them computationally expensive and impractical under limited data conditions.

1.3.3. Meta Learning with MAML model

Model Agnostic Meta Learning (MAML) algorithm is an optimization-based learning approach which helps in learning new tasks using small amount of data[14]. The core aspect is that it learns new tasks by drawing experience from prior related tasks. By learning an initialization for model parameters MAML ensures fine tuning on new model leads to fast convergence. The problem that could occur is the computational intensity due to the need of backpropagation through inner optimization steps[15].

One of the papers used. [15]MAML for face recognition tasks using small dataset Figure below, where the dataset is first divided into meta-training and meta-testing domains and each task set includes a support set and a query set. The support set is used to compute gradients and update the parameters during the first optimization step. While this is happening in first step, the query set will evaluate the loss, guiding the second optimization step through stochastic gradient decent. This is the two-stage gradient update which focus on enhancing the generalization ability of the model for new tasks. The result of the experiment showed a 92% accuracy.

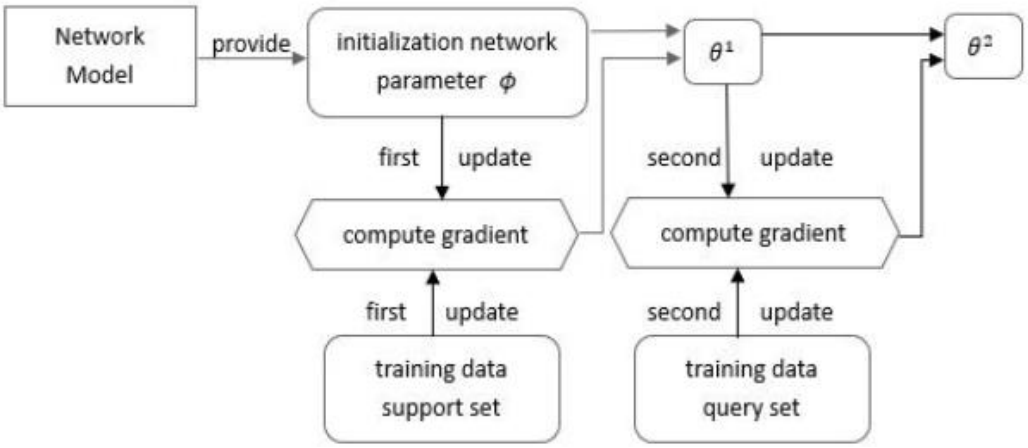

Figure 7. Represents the MAML Algorithm Process[15]

1.4. Semi Supervised Learning

The paper conducted different analysis of eight semi supervised learning models for facial expression recognition. The study was done using the dataset FER13, RAF-DB and AffectNet. The basic structure of the network is shown below.

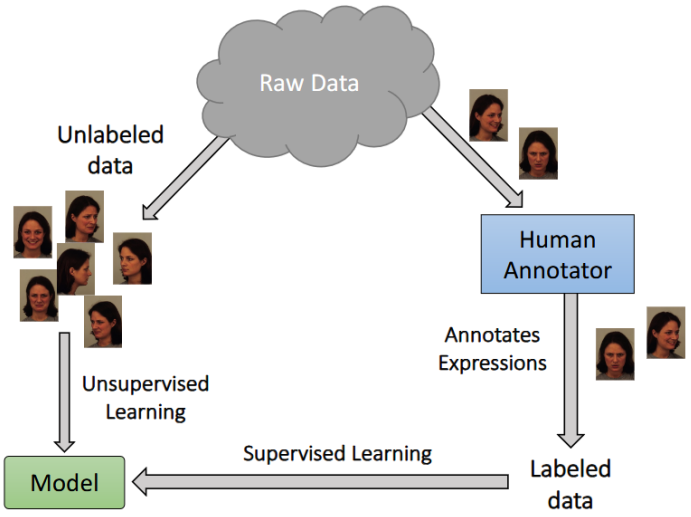

Figure 8. Represents the basic semi supervised model network[16]

The eight methods used are i-Model, Pseuro-label, Mean-Teacher, VAT, FixMatch, Re-FixMatch, UDA, and FixMatch which is also shown in the figure below. [16] [16]The experiment ended with a result where FixMatch outperformed all the other 8 instances and achieved second best in three other settings for FER13had 62%, RAF-DB had75% and AffectNet had around 51%.

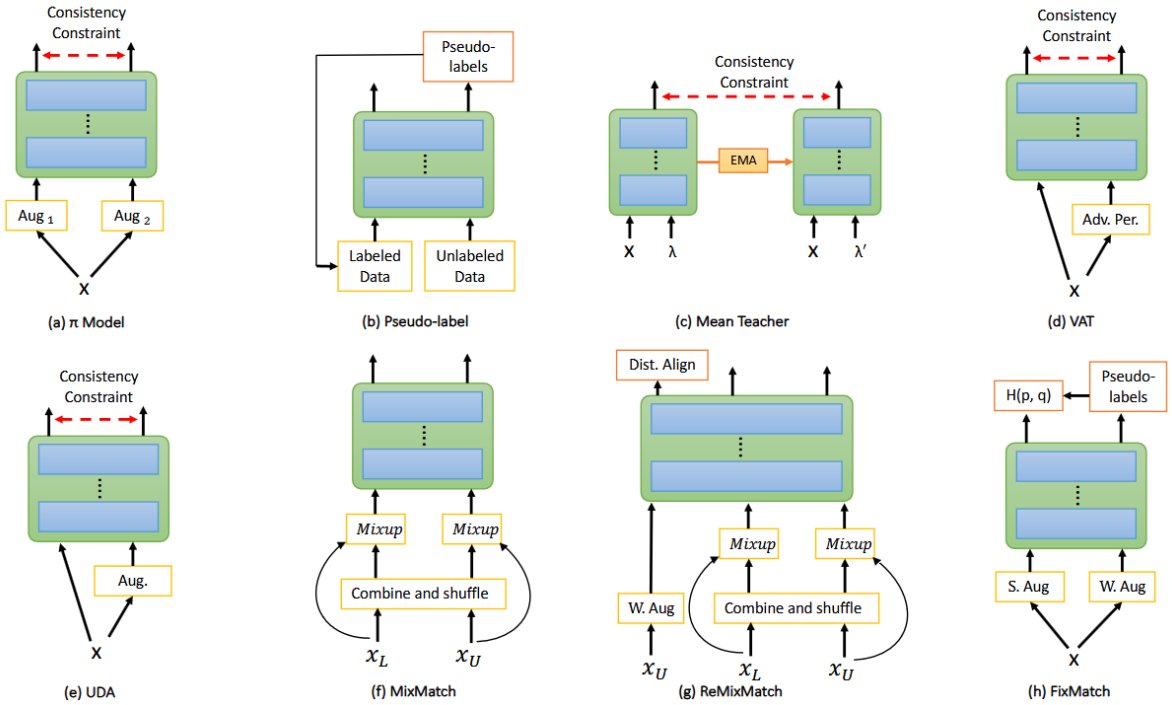

Figure 9. An overview of the different semi supervised methods [16]

1.5. Models used for FER with small dataset

This chapter will discuss the most popular models that are used for Facial expression detection with small dataset. Some of the models have already been introduced before but their architectures will be reviewed to get a better understanding.

1.5.1. AlexNet

The model first discussed in the first method was AlexNet mode CNN which was used as a two-step process for emotion recognition with small dataset. It is one of the shallow networks, which has 8 layers [1, 17].

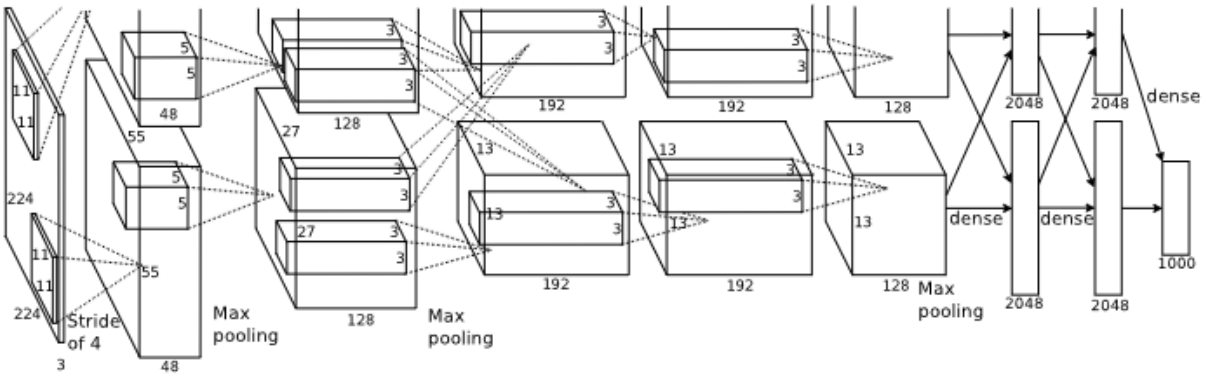

Figure 10. Represents AlexNet Architecture [17]

1.5.2. ResNet

RestNet50 was able to provide around 90% average in both experiments [5, 6]. The paper [18]where they replace the 2-layer block in the 34-layer net with 3 layer bottleneck block which results in 50 layer ResNet thereby increasing the accuracy [19]. The paper [20] [20]proposed an optimized ResNet architecture specifically for Fer, to demonstrate that targeted architecture; modification to standard residual network can improve the FER performance. However, ResNet based approach typically involve large parameter counts and require full model training, which limits their applicability in resource constrained and data scarce settings.

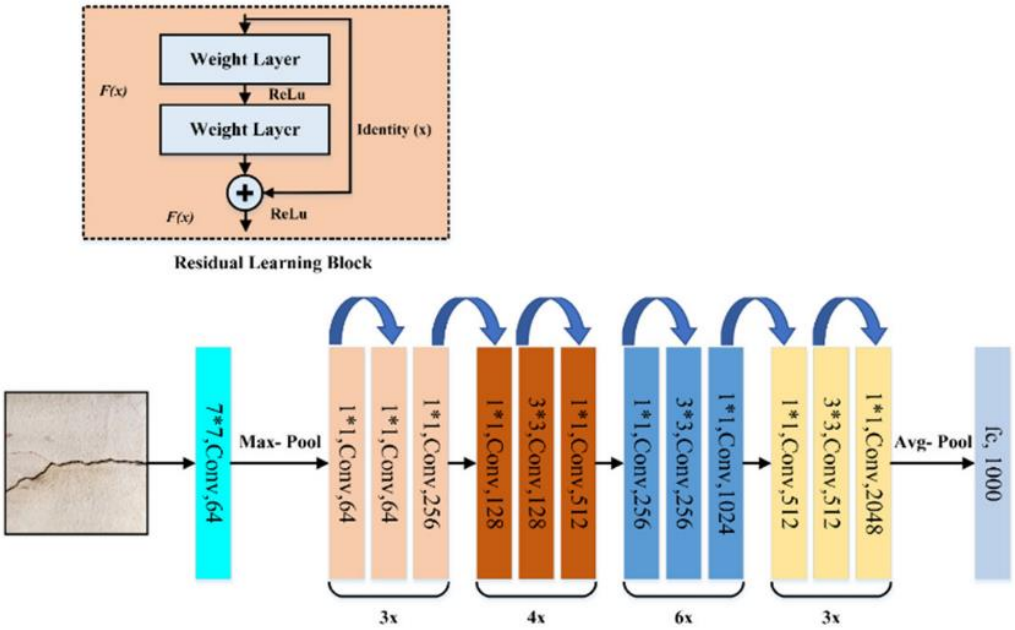

Figure 11. ResNet50 Architecture[19]

1.5.3. Datasets for FER

Selecting a good dataset for Facial Expression Recognition play a crucial role in advancing the research and practical application of emotion recognition system. These datasets are what is used as benchmark for training and evaluating the deep learning models. As the dataset is more diverse, it ensures better testing and development for FER system capable of handling different scenarios. Some of the widely used datasets are FER2013, RAF-DB, KDEF and JAFF[1, 4, 8].

FER2013 is a publicly available and popular dataset collected via the Google Search API, it features grayscale images resized to 48x48 pixels. The large training set of images around 28,000 and challenging in the wild nature has made it popular choice for emotion recognition tasks. But because of the grayscale images and fixed resolution it can limit its application in scenarios where high detail is required.

Figure 12. Example for FER Dataset[1]

RAF-DB which stands for The Real-World Affective Faces Dataset is another diverse, large-scale dataset which is designed to capture the facial expressions in an uncontrolled that is real world conditions. It has around 12,000 training images and higher resolution of 96x96 pixel [21].

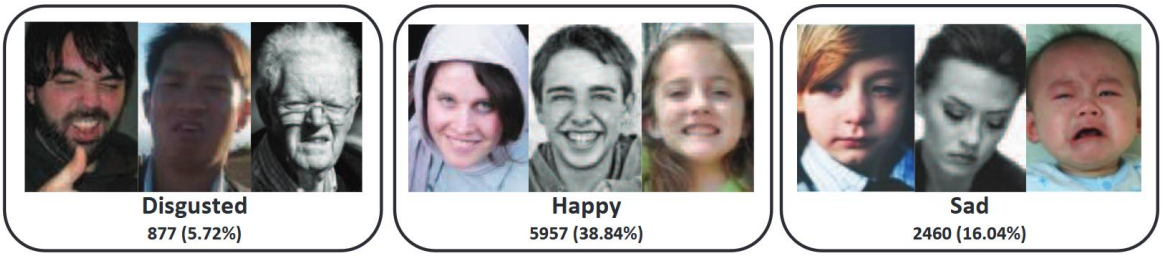

Figure 13. Example for RAF-DB Dataset [21]

KDEF or The Karolinska Directed Emotional Faces dataset is another dataset which is collected in controlled lab environment, having high resolution images of posed facial expressions. It is smaller compared to other datasets.

JAFFE, The Japanese Female Facial Expression dataset is a smaller but well-annotated datasets containing just over 213 grayscale images of posed facial expressions from 10 Japanese female subjects. The resolution is around 256x256 pixels and provides seven emotion categories and a consistent neutral background.

Figure 14. Example for JAFFE dataset [2]

1.5.4. The Challenges of Collecting Datasets

The papers have also discussed the different challenges they have faced like [4]discusses how the size of the labelled data was restricting the progress, as obtaining labelled data requires a tremendous amount of human effort, time and financial resources. Another issue faced when using active learning is the 'cold start' problem which occurs when the initial set of labelled data is small. On the other hand datasets like JAFFE and KDEF has another issue which is variability in facial expression which pose another challenge during training deep neural networks and if datasets are collected from real world situations like Emotiw from [1]and RAF-DB will contain noisy labels, occlusions and variations in lighting making them challenging for model generalization.

1.6. Techniques to Overcome Small Dataset Limitations

1.6.1. Data Preprocessing Methods

Preprocessing is the first step done to the images included[22], so that it can improve the performance of the FER system. Preprocessing is done so that the images will have more clarity, scaling contrast and additional enhancement process to improve the image quality so that the system will be able to identify the expressions easily. Some of the methods used were image alignment, bounding box definition, greyscale conversion, image resizing and normalization. During this process the images containing the face will be cropped where important facial components are included. Another method is normalization which helps minimise the external factors such as illumination changes and variations in face images which could affect the recognition process. Then comes localization, which helps in detecting the face, by spotting the size and locations of the face from image. Bounding boxes explain [7].

1.6.2. Data Augmentation Methods

When dealing with small dataset, it can result in overfitting, class imbalance where the number of images with happiness is more, or sadness is less and the difficulty with getting annotated data can be time consuming, to mitigate this, we perform different data augmentation methods. Data

augmentation is the manipulation of the data to provide more variable data from preexisting once. We can classify it to Geometric transformation and Intensity transformation [23]**.**

Geometric Transformation:

During geometric transformation some of the more popular or common methods used are techniques [24] are translation, where the image can be shifted horizontally or vertically to vary object position. Rotation in which rotating images by specific angles for various perspectives. Scaling, where the size of object is changed to simulate scenario with different scales. Flipping or mirroring the images horizontally or vertically to increase spatial diversity. Shearing or skewing is done to distort images diagonally to add variability to the orientation.

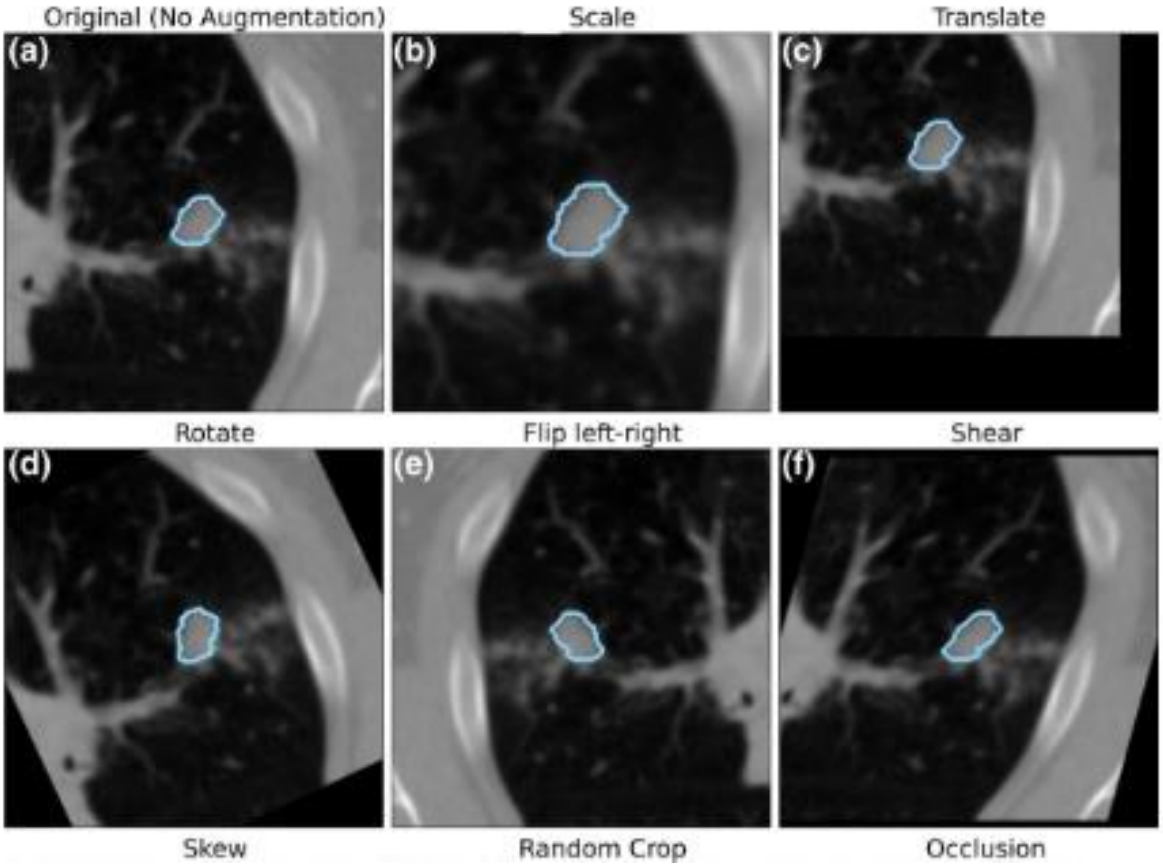

Figure 15. Example of commonly used basic data augmentation techniques with contour overlayed(blue) [23]

Intensity Transformation:

Intensity operation on the other hand works by manipulating the pixels within an image which can be by altering the brightness, contrast, noise levels or colour. The methods do play an important role in improving the model by creating variations that could simulate real-world conditions, like change in illumination or occlusion. Noise injection is a common method, where random or gaussian noise is added to an image to mimic the real-world imperfections like from the camera sensor noise. Brightness and contrast adjustments work by modifying the pixel intensity values, which in turn will

help the model generalize better. The cutout, Random Erasing and Grid Mask works by masking a part of an image to simulate natural obstructions. Cutout masks a fixed area, while Random Erasing changes its size and aspect ratio of the masked region dynamically. GridMask as the name implies applies a grid pattern to block parts of image in structured manner [23].

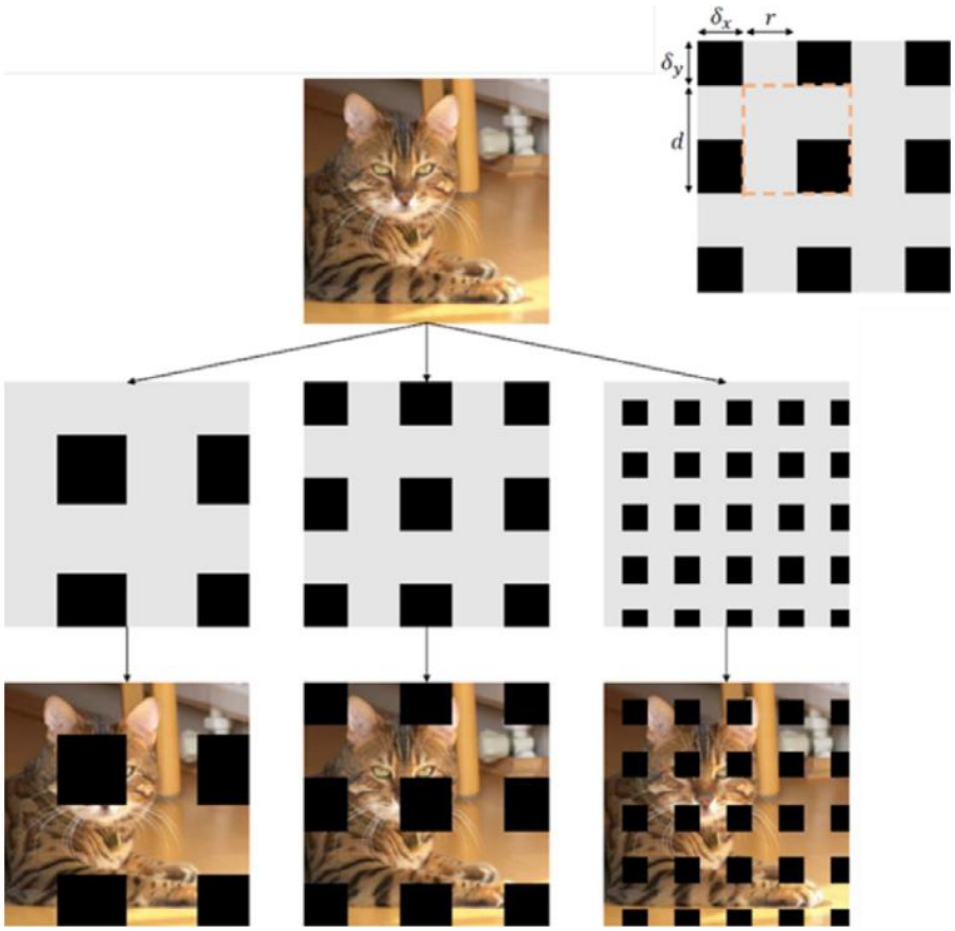

Figure 16. Example for intensity transformation GridMask [24]

1.7. Conclusion

1. From the literature analysis, multiple different facial expression techniques used recently were identified, which gave an idea on how to prepare the data sets and once prepared, which preprocessing methods must be applied and when to apply data augmentation methods, as there are situations which doesn't require preprocessing.
2. It also gave an idea on the different methods used for FER detection with small dataset and an idea on the architecture of the models used.
3. The analysis also found out about the different datasets that are used to conduct these experiments. The datasets most commonly used were FER2013, JAFFE, ck+ etc.
4. It also gave an insight into the different methods used for detecting FER.
5. The analysis also gave an idea on the different challenges that can be faced when using AI for facial expression detection, also about datasets and its collection.
6. Challenges that could be faced during training was also identified, and different methods to mitigate or reduce the issues were also discussed in this phase.

2. Proposed Methods for FER under Small-Data Constraints

2.1. Model Architecture and Hyperparameter Selection

This section will discuss some of the model architectures that were used in the following experiments. Different evaluation methods were used to capture the performance of the final trained model.

2.1.1. MobileNetV2

MobileNetV2 is a lightweight model which is designed to work well on mobile phones and small devices, while also being able to perform [25]tasks like image classifications. The model was primarily trained on ImageNet dataset, which is a large-scale image dataset used for visual recognition tasks. The main strength of the model is its use of depthwise separable convolutions and inverted residual blocks. This design features significantly reduce the number of parameters and computations, while still preserving the model's ability to learn complex features. The model has an input layer which takes in 224x224 RGB images and then applies a regular 3x3 convolution to detect the simple patterns like edges. The Inverted Residual Block are the main building blocks of MobileNetv2, where each block does three things, expands the image features using a 1x1 convolution, filters them using a 3x3 depthwise convolution, and finally shrinks them back down with another 1x1 convolution. Then it has pointwise Convolutions, which are 1x1 convolutions used to mix features and reduce the number of channels, thereby making the model faster. The final layers averages everything into a small vector and then it uses the fully connected layer. The figure 17 represents the building block of mobilenetv2. On the left side, the inverted residual block with shortcut connections or skip connections is present.

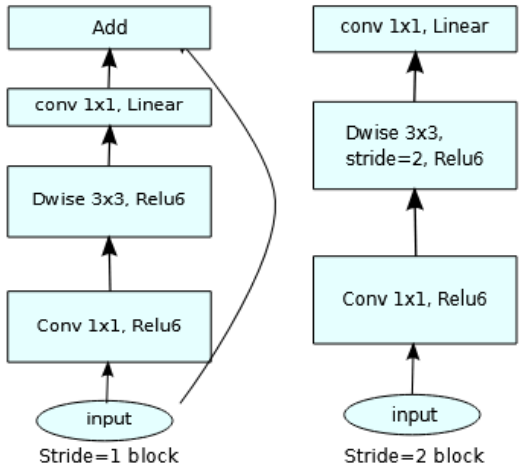

Figure 17. MobileNetV2 [25]

2.1.2. Hyperparameter Selection

The hyperparameters used in this work were selected based on commonly reported values in the literature for transfer learning and refined empirically using validation performance. Learning rates were chosen from the range 10^{-3} to 10^{-5} to balance convergence speed and training stability. Batch sizes were constrained by data size and GPU memory limitations.

2.2. Baseline Transfer Learning Models

2.2.1. MobileNetV2

For the experiment, transfer learning using a pretrained model called MobileNetV2 was applied on the JAFFE dataset, which was discussed earlier. The experiment was done by first preprocessing the images as the dataset is small and requires data augmentation to reduce or prevent over fitting. From the methods of preprocessing discussed before, the ones implemented in this experiment are converting the single grayscale to 3- channel RGB format as MobileNetV2 expects a 3-channel input. Then applied rotation of ±20 degrees, as it simulates the variation of head tilt. Random resized crop was applied next, where the image was resized to 256 pixels and a scale of 0.8 and 1.0, a random flip transformation was applied with a probability of 0.5. Normalization was not applied in this method as the results obtained were not great.

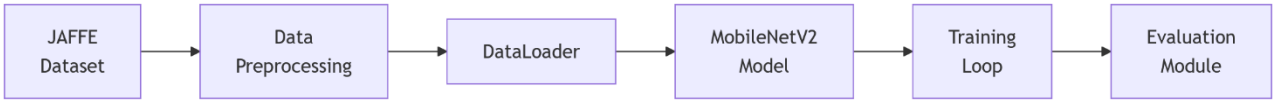

Figure 18. Transfer Learning Architecture

To train the MobileNetv2 effectively on the JAFFE dataset, the dataset was split into training and testing as a sperate testing dataset was not available for the given dataset. This ensures that each model class was represented in both sets. In this split almost 90% of the data was used to train the model while rest of the 10% was used to test the model. As the size of the dataset is significantly small, data augmentation was applied to the training dataset in order to reduce the overfitting and simulate the real world variations. The model was trained on 20 epochs using the Adam optimizer with a learning rate of 0.0001. The classification loss was calculated using crossentropy function. The memory efficiency and training stability was maintained by using a batch size of 32. Once the training was done it was evaluated using the test set which was reserved before.

The last step to achieve this was modifying the final classification layer of the model to suite the 7 emotion classes which is based on the [2].

Table 1. Parameter used for normal pretraining with mobilenet v2

Parameters	**Value**
Modified Layer	Linear(1280, 1000) to Linear(1280, 7)
Loss Function	CrossEntropyLoss
Optimizer	Adam
Learning rate	0.001
Epochs	20
Batch Size	32

2.2.2. ResNet-50

Here the transfer learning method was applied using the model ResNet50. For this method basic preprocessing techniques where applied such as image resize and normalization. The dataset used to train the model is FER2013 dataset, which has both training and test dataset.

Architecture Components

The model begins with an input layer where ResNet50 takes in input images of size 224x224x3 which is pre-processed. The convolution and pooling layer has a 7x7 convolution layer with a stride 2 and a max pooling layer which is followed by four residual blocks, where block1 has 3 bottleneck layers, block 2 has four bottleneck layers block 3 has six bottleneck layers and block 4 has 3 bottleneck layers. Each block contains a convolutional layer along with a skip connection. Originally the resnet ends with 2048 with 1000 classes, which is for ImageNet, but in this situation, the final layer is replaced with 7 to match the seven emotion classes. The final output layer outputs a vector of 7 logits, with one for each emotion class.

Table 2. Hyperparameters used for training ResNet50

Parameters	**Value**
Modified Layer	Linear(2048, 1000) → Linear(2048, 7)
Loss Function	CrossEntropyLoss
Optimizer	Adam
Learning Rate	0.001
Epochs	10
Batch Size	32

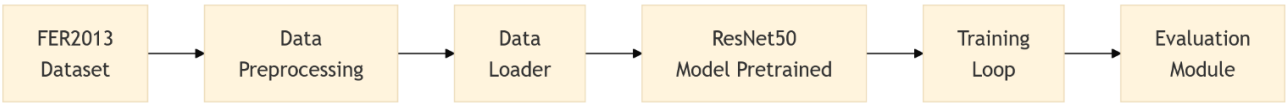

Figure 19. Transfer Learning ResNet50 Architecture

2.3. Advanced Transfer Learning Strategies

2.3.1. Two-Stage Transfer Learning with MobileNetV2

This method was discussed in paper[1], where the model was first trained on a bigger dataset and then it was later trained again on a smaller dataset. The reason why this method was opted is because small dataset is prone to overfitting. Hence first a CNN pretrained on ImageNet is fine-tuned on a smaller specific class-based dataset like FER2013 where the class is emotions or facial expressions. Since the dataset and model used in the paper was unavailable, the same method was approached with another model called MobileNetV2, the model was opted as it is lightweight and is pretrained on ImageNet dataset.

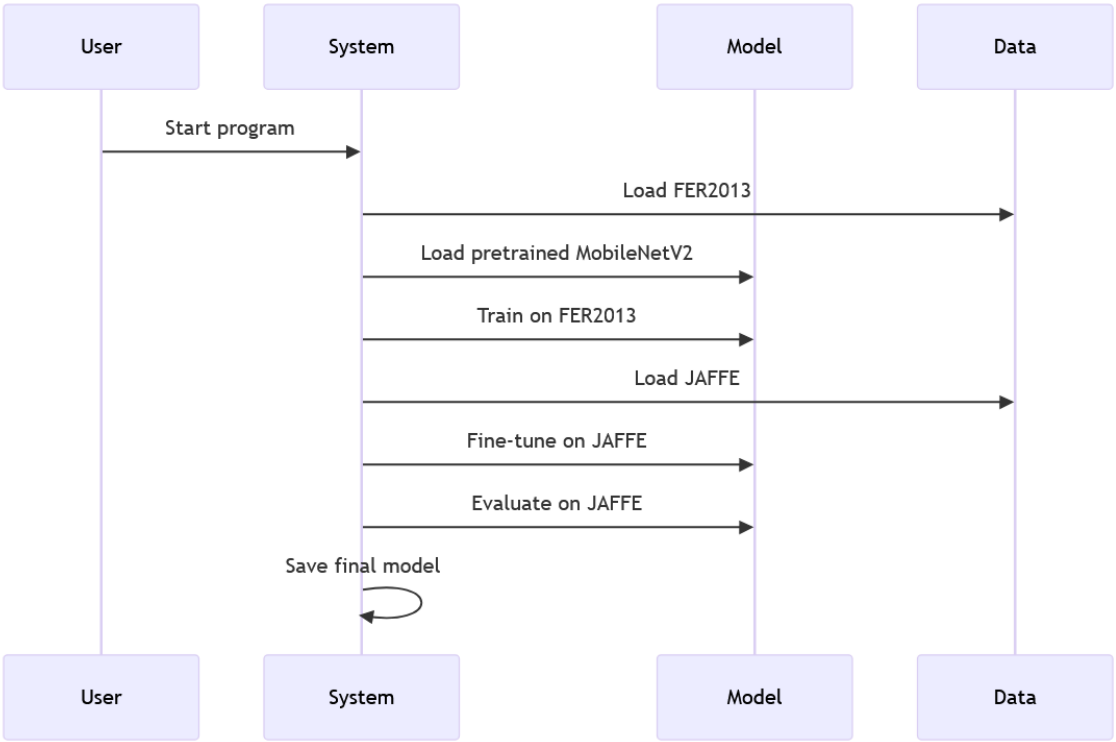

Figure 20. Architecture of Two stage transfer learning with MobileNetV2

Model Architecture

So before giving to the model, the dataset used here are FER2013 and JAFFE dataset as the dataset mentioned in the paper were unavailable. First stage of data preprocessing was done on FER2013 dataset, where the transforms applied are, converting the grayscale to three channels, resize the images to 224x224 and normalization. For the second stage, the dataset is split into training, and validation sets with random split.

The model was trained for 15 epochs in total, 10 epochs for the FER2013 dataset and 5 epochs for the JAFFE dataset, which is the stage two fine-tuning, and the layer was modified as done previously, so that it can accept the seven classes. The remaining values were applied similar to the paper to observe if the output could be closer to the original paper.

The following settings were kept the same, such as same optimizer as SGD and weight decay = 0.0005 with momentum = 0.9. Weight decay helps in keeping the model weights small and generalizable while momentum on the other hand helps speed up the convergence and stabilize the training. The model was evaluated using the matrix evaluation to see if the images are generalized properly by the model and the f1 score and other matrices were used to see if the model suffered any overfitting.

Table 3. Parameter for training MobileNetV2

Parameters	**Value**
Modified Layer	Linear(1280,1000) -> Linear (1280,7)
Loss Function	CrossEntropyLoss
Optimizer	SGD
Learning Rate	0.001 (Stage 1), 0.0001 (Stage 2)
Epochs	10 (FER2013) + 5 (JAFFE) = 15 total
Batch Size	64 (FER2013), 32 (JAFFE)
Momentum	0.9
Weight Decay	0.0005

2.3.2. Two stage pretraining with AlexNet

The same two stage fine tuning was applied using AlexNet to see how the result will fair against mobilenetV2. The setup is similar to the previous model, where for the first stage FER2013 dataset is used, while for the second stage the dataset used will be JAFFE. Image preprocessing is done same as the previous experiment. The hyperparameter applied can be seen from the table below.

Table 4. Parameter for training AlexNet

Parameters	**Value**
Modified Layer	Linear(4096, 1000) → Linear(4096, 7)
Loss Function	CrossEntropyLoss
Optimizer	SGD
Learning Rate	0.001 (Stage 1), 0.0001 (Stage 2)
Epochs	15 (FER2013) + 10 (JAFFE) = 25 total
Batch Size	32 (both FER2013 and JAFFE)
Momentum	0.9
Weight Decay	0.0005

2.4. Proposed Efficient PathMobileNet Architecture

The proposed methodology is a combination of using ideas inspired by neural architecture search (NAS) [13]principle and path-based network selection [12]and focuses on identifying an optimal sub-network within a fixed MobileNetv2 backbone for facial expression recognition.

In this proposed approach, instead of directly adopting the full pathNet formulation, this work introduces simplified and computationally efficient architecture search strategy tailored specifically for facial expression recognition (FER) under limited data conditions.

Unlike PathNet, which is designed for multitask and meta-transfer learning scenarios, the proposed approach focuses on a single-task FER setting using a fixed pretrained backbone. MobileNetV2 is selected as the base architecture due to its lightweight design and suitability for resource-constrained environments.

2.4.1. Problem Formulation

Let MobileNetV2 be composed of a sequence of 17 inverted residual blocks. Rather than utilizing the full network, a subnetwork is formed by selecting the first k blocks. where $k \in \{3 \ldots\ldots 17\}$. The path always begins at block 1, only the endpoint is varied during the search. Given a limited labelled dataset, the objective is to identify the subnetwork that provides the best trade-off between model capacity and generalization, evaluated using a lightweight logistic regression.

The problem can be formulated by selecting the subnetwork that maximizes the classification performance under limited data constraint. This enables efficient exploration of model capacity, without requiring full network training.

2.4.2. Architecture Design

In the proposed framework, MobileNetV2 is decomposed into its constituent inverted residual blocks, enabling structure exploration of sub-networks. The initial convolutional layer and the final convolution and pooling layers are kept fixed, as they are responsible for low-level feature extraction and high-level semantic representation. The intermediated inverted residual blocks are what form the search space from which candidates for the sub-network, are constructed in a sequential path of consecutive blocks, always starting from the first block and extending to a variable depth (blocks 1-k). This constraint ensures a consistent feature hierarchy while allowing systematic exploration of network depth.

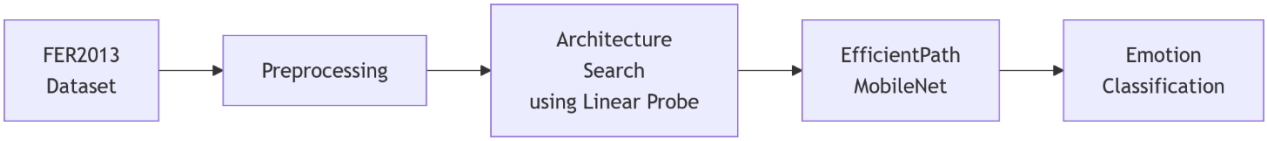

Figure 21. Architecture Overview

Only the selected candidate path will be instantiated during the model construction, while all unused blocks are completely removed from the network. To ensure compatibility between blocks with different channels dimensions, lightweight input and output adaptation layers based on 1x1 convolutions are introduced when necessary. This design enables genuine parameter reduction while preserving the structural integrity of the backbone.

Each inverted residual blocks follows a three-layer structure, a pointwise convolution that expands the channel dimensions, a depth wise convolution that processes spatial features and a second

pointwise convolution that projects back to a lower dimension. A skip connection is added when the input and output dimensions match, which help preserve gradient flow during training.

The architecture is sequential in nature, that is features become progressively more abstract as data passes through each block. This property makes it naturally suited for structured subnetwork selection, as candidate subnetworks can be formed by simply truncating the network at block k while retaining a consistent feature hierarchy.

2.4.3. Search Strategy

The search space for this strategy is made up of a sequence of inverted residual blocks, where each candidate subnetwork includes blocks from layer 1 up to layer k, with $k \in \{3 \ldots . 17\}$. This design will allow exploration of different network depths while maintaining structural consistency.

To efficiently evaluate each candidate, [26]a linear probing method is used. The convolutional backbone is kept frozen, and features are extracted once for all the candidates and cached to reduce the computation. A logistic regression classifier is then trained on these features to estimate the classification accuracy of each candidate subnetwork. This approach will provide a fast and reliable estimate of each architecture without the computational cost of full end to end training. Linear probing work well as a quick test and can predict how well a model will perform on real task, so it is useful for evaluating the architecture when resources are limited.

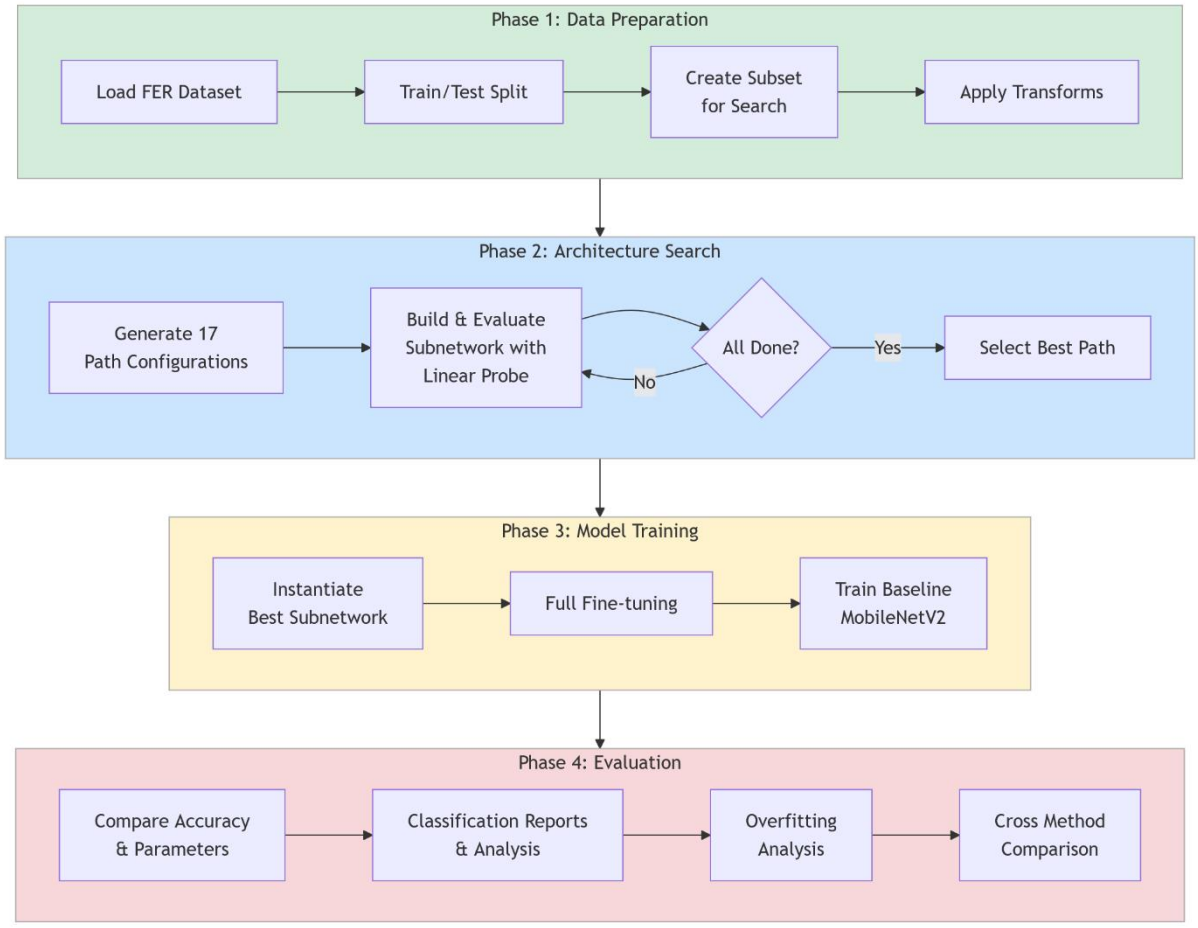

Figure 22. Proposed Model Detailed Architecture

For FER,2013 all candidates are evaluated using two non-overlapping 15% subsets from the training data alone. One for fitting the classifier and one for evaluation, without data augmentation. The search is performed using a single fixed seed to ensure reproducible results.

For CK+, the search subset is constructed using a subject independent split, where subjects are assigned to either the search training or search evaluation, so that no subject appears in both. This is done to prevent data leakage during the search phase.

Compared to the Neural Architecture Search approaches, the proposed method provides a significantly more efficient alternative, making it more suitable for low data and resource constrained settings.

2.4.4. Training Strategy

To evaluate the performance of the proposed subnetwork under both standard and low data conditions, experiments are conducted both in two datasets, FER 2013 and CK+.

For FER2013 dataset, a limited data regime is simulated by training the models on different fractions of the training set, which are at 5%, 10%, 20% and 100% of the available data. For each fraction, a subset of emotion is sampled based on the random seed. An internal 80/20 train/validation split is applied within the sampled subset, where the validation split is used for early stopping and the final accuracy is evaluated on the held-out test. Each experiment is repeated across five different seeds, resulting in a 25 different training configuration in total.

The comparison is done between two models, the full MobileNetv2 and the proposed parameter efficient subnetwork. Both models are initialized with ImageNet pretrained weight and trained with identical settings to ensure a fair comparison of the models. The models use Adam Optimizer with a learning rate of 0.0001, early stopping with a patience of 7 epochs, a maximum epoch of 50, labelled smoothing of 0.1 and weighted Random sampler to handle the class imbalance. For both the models, only the initial convolutional block is frozen during training. During training standard data augmentation techniques are applied, such as random horizontal flipping and rotation. At the end both models are tested on the full test dataset without augmentation.

For the CK+ dataset, a subject independent five-fold cross validation protocol is used in order to prevent data leakage. Within each fold, the non-test subjects are further divided at the subject level into training set (85%) and an internal validation set (15%). The validation set is used for early stopping and checkpoint selection, while the test fold is kept completely unseen until final evaluation. This three-way subject level split ensures no subject images appear in more than one partition, preventing data leakage. The cross validation is repeated across three random seeds, resulting in 15 evaluations per model.

Final Model Selection and Training

After the architecture search phase, the best performing candidate's architecture is selected based on validation accuracy obtained during linear probing. The selected architecture is then fully trained end to end on the complete FER2013 training set using standard cross entropy loss.

3. Experimental Results and Discussion

3.1. Experimental Protocol

This section will provide a consolidated overview of the experimental procedures followed in this work. Detailed description of datasets, architecture and results are presented in the subsequent sections.

The datasets used are FER2013, CK+ and JAFFE, with its training and testing split for the subnetwork search and limited data experiments. CK+ was used for the subject independent cross validation of the proposed subnetwork using a five-fold method at the subject level to prevent any data leakage. Jaffe was used mainly for the initial baseline transfer learning experiment and was randomly split into training and testing subsets as it lacked a predefined split.

All experiments followed a consistent pipeline consisting of data preprocessing, optional data augmentation, model training and evaluation. Preprocessing and augmentation methods were selected based on the requirements of each pretrained architectures as described in the methodology section. For FER 2013 and CK+, grayscale images were converted into three channel RGB by duplicating the single channel, and all images were resized to 224x224 pixels to match the ImageNet input requirements of the pretrained backbones.

Models were trained using either the standard transfer learning, two stage fine-tuning or the proposed subnetwork selection method. For the subnetwork search, features were extracted once from the frozen backbone and evaluated using logistic regression on a fixed subset of the training data, with the test set kept completely unseen during the search phase. For full training experiments, early stopping with a patience of seven epochs were applied to prevent overfitting and performance was evaluated using a held-out test set. Both models used identical training settings including the Adam optimizer and a learning rate of 0.0001, label smoothing of 0.1 and a weighted random sampler to address the class imbalance.

All the experiments were conducted on a single machine setup using a consumer grade GPU, specifically an RTX 4060. Due to computational constraints, the architecture search was performed using linear probing on a reduced dataset subset, while full model training was reserved for the selected architecture only.

3.2. Dataset Overview

The datasets that were used for this study, are FER 2013, CK+ and JAFFE dataset. These are opensource or public datasets. The JAFFE dataset consists of Japanese women facial expression and FER consist of human emotion dataset. Both datasets have seven expressions. The FER 2013 dataset consist of 35,887 grayscale images, with size of 48x48 pixels. The images depict facial expression in an unconstrained environment. The dataset is split into two sets, a training set and a testing set. The dataset's variability in lighting, pose and background makes it valuable for training models with real-

world conditions. FER-2013 is available through Kaggle platform and is released under a public domain or research license.

The second dataset used is the JAFFE mentioned before. It is a small dataset consisting of just 213 grayscale images of size 256x256 pixels of 10 Japanese female students. The FER dataset offers large-scale, real-world data while the JAFFE provides, small-scale high-resolution data. JAFFE can be accessed via Zenodo and is intended for academic research purposes under a standard research license and the papers mentioned in the website has to be cited as reference [27, 28].

The third dataset used is the Extended Cohn-Kanade dataset also known as the CK+. It is also a lab-controlled benchmark dataset, which consist of image sequences from 123 different subjects across seven emotion categories, including anger, contempt, disgust, fear, happiness, sadness and surprise. Unlike FER2013 which contains images from unconstrained real-world environments, CK+ provides controlled conditions with consistent lighting and frontal face alignment. The dataset contains 593 sequences in total and is used widely in subject independent evaluation in FER research. In this work, a subject independent five-fold cross validation protocol is applied so that no subject will appear in both training and test sets simultaneously, which helps prevent data leakage during evaluation.

Different datasets were used at different stages of this study based on their characteristics and suitability for each experiment objective. JAFFE was used for the initial baseline transfer learning experiments due to its small, controlled nature, which is appropriate for evaluating basic finetuning under limited data. FER2013 was used for the subnetwork search and limited data experiments due to its sufficient size for simulating multiple data fraction. CK+ dataset was used for the subject independent cross validation of the subnetwork, as its subject level structure enables an evaluation protocol that prevents identity leakage. This multi dataset approach allows the proposed method to be assessed under both controlled and uncontrolled conditions.

3.3. Pre-processing steps applied

The dataset was used in different models and different techniques, so the preprocessing method applied for the different task might vary as some performed better with one type of preprocessing, while the same preprocessing gave bad result for another model or method used. But a general preprocessing steps used are listed as follows:

- **Converting Images to RGB**

Some models like AlexNet are pretrained on RGB images and these models won't accept grayscale images. So, the grayscale images were converted to 3-channel RGB by duplicating the single grayscale channel.

- **Resize**

In this method the images were resized to 256x256 or 224x224 pixels for models likes ResNet. This is because these models were trained on ImageNet dataset, where the input images are resized to 224x224 before training.

- **Randomized Crops**

This method is used to randomly crop and resize a portion of image to 256x256. It simulates the zooming and slight translation.

- **Horizontal Flip**

This method will randomly flip the image horizontally with a probability of 0.5.

- **Random Rotation**

Rotate the image randomly with ±10 degrees or ±20 degrees. It helps simulate head tilts or variation in pose.

- **Normalize**

It normalizes the pixels values from range [-1,1] using custom mean and std. For some method as the paper which was used to design the same architecture said that in some situations normalization made the situation worse hence was ignored.

3.4. Evaluation Methods applied

This section describes the various classification evaluation techniques that is used to evaluate the models and its performance. The classification assessment or evaluation methods are used to see how well a model is making its predictions. Some models will give a yes or no answer like a decision tree, while others will give a probability like Naïve Bayes. But no matter how they work, to assess how good their predictions are the following methods were used: [29]

1. Precision

It tells us how many of the items that the model said were correct is actually correct, that is it shows the proportion of true positive predictions among all positive predictions made by the model.

2. Recall

It tells us how many of the actual correct items the model was able to find. In other words, it looks at how many of the actual positive cases the model was able to identify.

3. F1-Score

It balances the precision and recall, can be used in situations where some classes will have fewer samples.

4. Support

It shows all the examples that are present in each class.

5. Accuracy

As the name implies tells us how many of the total predictions that were made was actually right out of all the predictions.

6. Micro Average

It treats all the classes equally and it adds everything up before calculating the precision, recall or f1. It's really good in situations where all the predictions are equally important. It calculates all the metrices by counting the total true positives, false positives and false negatives.

7. Confusion Matrix

Confusion Matrix basically is a table, which is mainly used to describe the performance of a classifications model. The matrix shows how many of the predictions that the model made is correct and which ones it made a mistake. This is done by comparing it with the class labels to the labels that were predicted. In a binary classification, the confusion matrix has four parts which are

True Predictions (TP): The model correctly predicted the positive class to which the sample belongs to. Eg: the sample belongs to class x and the model correctly predicted class x.

True Negative (TN): The model correctly predicted the negative class to which the sample belongs to. Eg: the sample does not belong to class x and the model correctly did not predict the class x.

False Positive (FP) : These predictions where incorrectly predicted as positive, that is the sample does not belong to class x, but the model predicted that the sample belongs to class x.

False Negative (FN): Here the model incorrectly predicted as negative, i.e. the sample belongs to class x but the model predicted a different class.

The rows represent the actual class while the columns represent the predicted classes. The diagonal values of the matrix represent the predictions per class which are actually correct, while the ones off the diagonal are the values which shows misclassification indicating that the model is a bit confused.

3.5. Baseline Model Performance

3.5.1. Observation Result for MobilenetV2

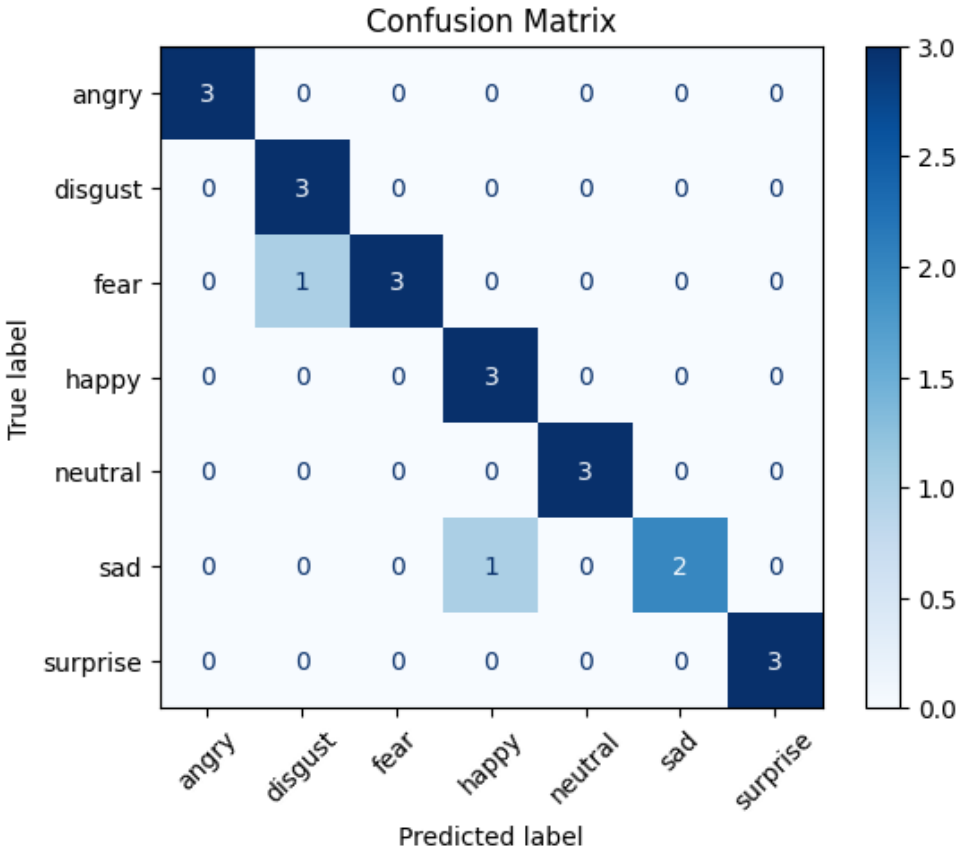

Figure 23. Confusion matrix of MobileNetV2

```
Epoch 1/20, Loss: 1.7711, Accuracy: 30.89%
Epoch 2/20, Loss: 1.1673, Accuracy: 54.97%
Epoch 3/20, Loss: 0.8607, Accuracy: 67.54%
Epoch 4/20, Loss: 0.6093, Accuracy: 81.15%
Epoch 5/20, Loss: 0.5679, Accuracy: 80.10%
Epoch 6/20, Loss: 0.4785, Accuracy: 81.15%
Epoch 7/20, Loss: 0.4004, Accuracy: 85.34%
Epoch 8/20, Loss: 0.3215, Accuracy: 89.01%
Epoch 9/20, Loss: 0.2318, Accuracy: 91.10%
Epoch 10/20, Loss: 0.3251, Accuracy: 86.91%
Epoch 11/20, Loss: 0.2725, Accuracy: 90.58%
Epoch 12/20, Loss: 0.4199, Accuracy: 81.15%
Epoch 13/20, Loss: 0.3095, Accuracy: 87.43%
Epoch 14/20, Loss: 0.1852, Accuracy: 94.76%
Epoch 15/20, Loss: 0.2101, Accuracy: 91.62%
Epoch 16/20, Loss: 0.2802, Accuracy: 91.10%
Epoch 17/20, Loss: 0.2101, Accuracy: 92.67%
Epoch 18/20, Loss: 0.1787, Accuracy: 93.72%
Epoch 19/20, Loss: 0.2375, Accuracy: 92.15%
Epoch 20/20, Loss: 0.1387, Accuracy: 94.24%
```

Figure 24. Training Loss and Accuracy

Evaluation and Result

During training, the model accuracy increased from 30.89% to 94.24% in 20 epoch, along with that the training loss also decreased from 1.77 to 0.14 which can be noticed from the figure 24. After training, the model was evaluated on a test set, which achieved an accuracy of 77.27% which shows good generalization. Based on the confusion matrix, it indicates the followings:

1. The model perfectly classified six out of seven emotion categories.
2. Angry, Disgust, Happy, Neutral and surprise each have 3 correct predictions and zero misclassification.
3. Fear had three correct but also one false positive from the disgust class
4. Sad was the only class with more noticeable confusion.
5. From this it can be observed that the model generalizes well for most classes, especially, happy, surprise and neutral.

3.5.2. Observation and Result ResNet 50

Table 5. Accuracy and classification report on ResNet50 transfer learning

	Precision	**Recall**	**F1-score**	**support**
Angry	0.55	0.62	0.59	958
Disgust	0.62	0.61	0.62	111
Fear	0.52	0.49	0.50	124
Happy	0.88	0.83	0.85	1774
neutral	0.61	0.62	0.61	1233
Sad	0.53	0.55	0.54	1247
surprised	0.81	0.79	0.80	831

Accuracy			0.66	7178
Micro average	0.65	0.64	0.64	7178
Weighted average	0.67	0.66	0.66	7178

Figure 25. Confusion matrix of ResNet50

Evaluation and Result

From the classification report from the table, it can be observed that, the model shows an accuracy of 66%. It has a weighted average f1 score of 0.66, which shows that the model performs reasonably well across classes but has room for improvement. The best performing classes from the report are happy and surprised classes. Which indicates a strong model confidence and consistency in predicting these emotions. The low performing classes were fear and angry with and f1 score of 0.50 and 0.59.

From the confusion matrix happy and surprise had the correct predictions, with happy being the highest of all classes. Sad often got confused with angry and fear, fear got predicted as sad and angry frequently and angry has a high confusion with sad and fear.

3.6. Two Stage Model Performance

3.6.1. Observation and Result for Two Stage MobileNetV2

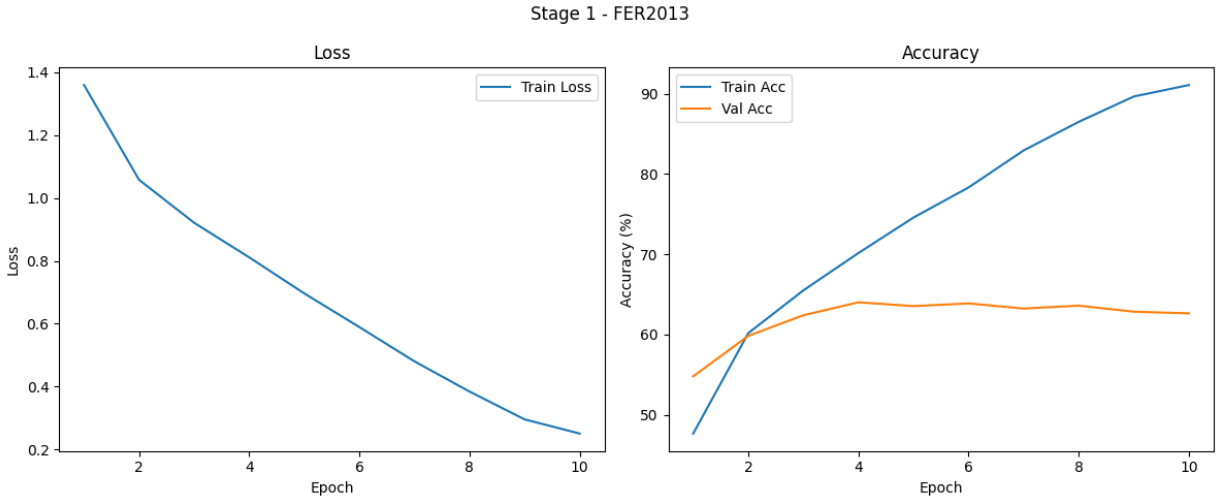

Figure 26. Stage 1 Fer 2013

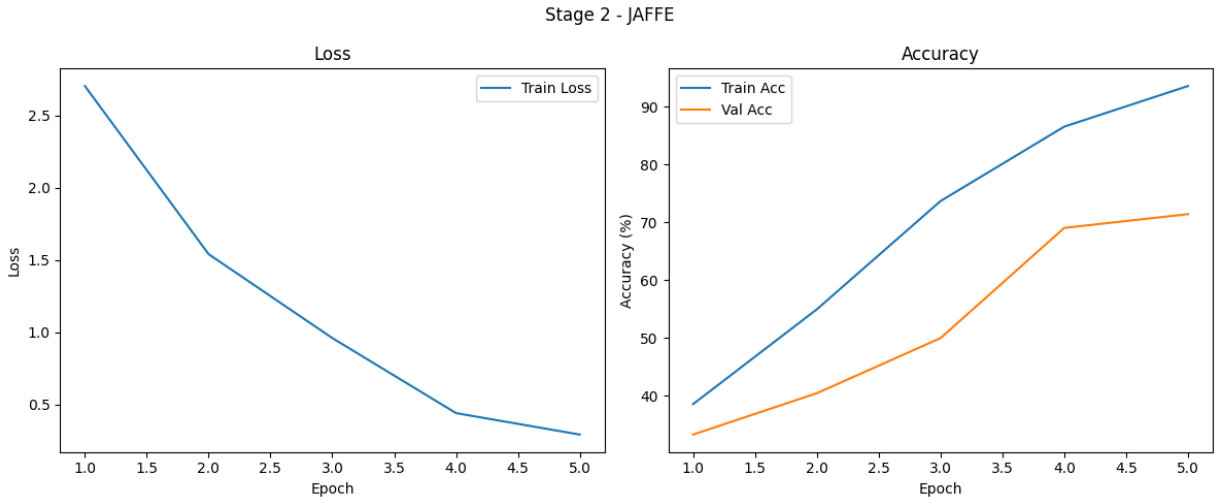

Figure 27. Stage 2 JAFFE

Table 6. Accuracy and classification report on Two stage- transfer learning process

	Precision	**Recall**	**F1-score**	**support**
Angry	0.50	0.80	0.62	5
Disgust	0.50	0.25	0.33	4
Fear	0.80	0.67	0.73	6
Happy	1.00	0.83	0.91	6
neutral	0.50	1.00	0.67	3
Sad	0.67	0.60	0.63	10
surprised	1.00	0.88	0.93	8

Accuracy			0.71	42

Micro average	0.71	0.72	0.69	42
Weighted average	0.71	0.71	0.71	42

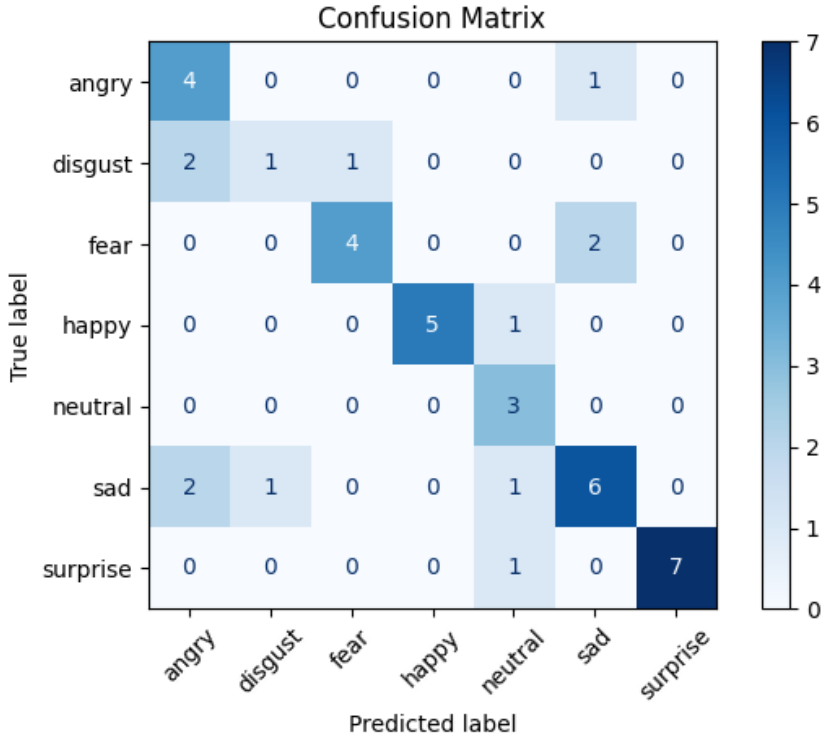

Figure 28. Confusion Matrix for two stage transfer learning

Evaluation and Result

From the Fig 26 in stage 1 it can be observed that the training loss steadily decreases over 10 epochs, which means that the model is learning effectively from the FER2013 dataset.

The accuracy curve on the other hand indicates that, the training accuracy increases consistently, reaching above 90%, but the validation accuracy plateaus around 63-65% after 4 epochs and showed fluctuations. Which suggests that the model is experiencing overfitting.

Fig 27 on the other hand which represents the stage 2 of the transfer learning process, where the JAFFE model was used. Here the training loss drops sharply, especially after the first two epochs, indicating fast learning due to the transfer learning. By the 5th epoch, the loss is very low, which indicates a strong convergence.

The accuracy curve in stage 2 quickly rose from 40% to over 90%, which means that the model was adapting well to the new dataset. The major difference is that the validation accuracy also showed significant improvement, when compared to stage one, where here it went from 33% to over 70%, indicating good generalization.

Unlike stage one the improvement in the validation accuracy, indicates less overfitting in this phase.

From Table 6 , it is observed that the accuracy is 71.43%. The precision, recall and F1-score is around .70 in average which indicates balanced performance across the classes.

From the inspection of the individual classes, it can be observed that the model performed well on classes like “fear”, “happy”, and “surprise”. Sad had an f1 score of 0.31 and disgust had an f1 score of 0.57 which indicate poor performance.

The confusion matrix from fig 28 provides a detailed view on how well the model performed against each class. The fig indicate that Disgust was heavily misclassified, primarily predicted as anger and fear. Sad also showed confusion, with samples misclassified as angry and disgust. Fear had some confusion with sad. Neutral and surprise on the other hand were perfectly or near perfectly classified, with neutral achieving 100% recall and surprise identifying all seven samples.

3.6.2. Evaluation and Report for Two Stage AlexNet:

Table 7. Accuracy and classification report on Two stage-transfer learning process AlexNet

	Precision	**Recall**	**F1-score**	**support**
Angry	0.25	0.40	0.31	5
Disgust	0.20	0.25	0.22	4
Fear	1.00	0.50	0.67	6
Happy	0.75	0.50	0.60	6
neutral	0.33	1.00	0.50	3
Sad	1.00	0.30	0.46	10
surprised	0.80	1.00	0.89	8
Accuracy			0.55	42
Micro average	0.62	0.56	0.52	42
Weighted average	0.70	0.55	0.55	42

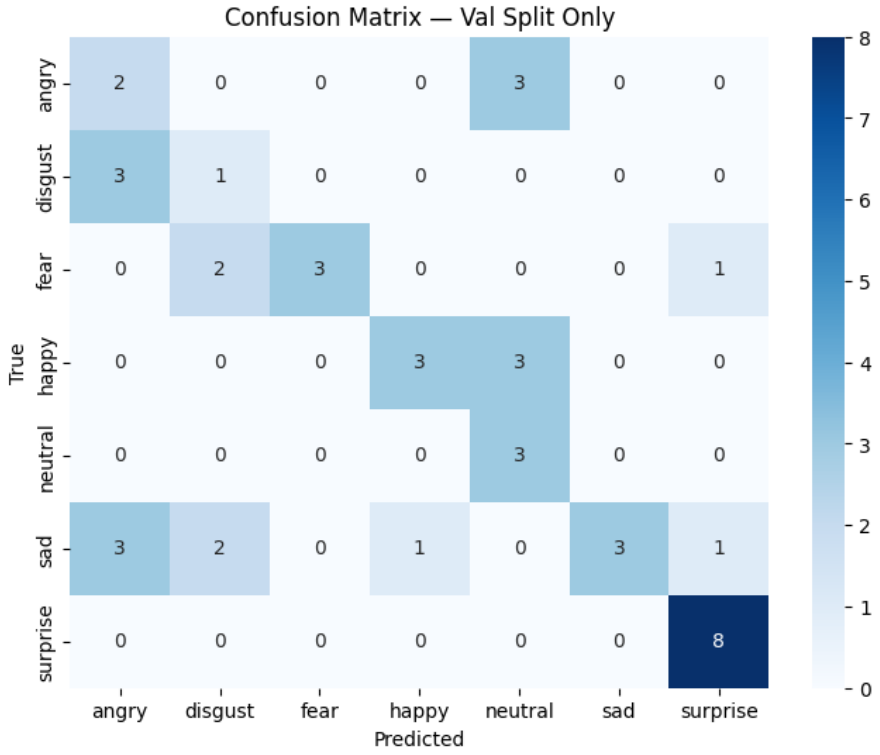

Figure 29. Confusion matrix of AlexNet

Evaluation and Result

The table 7 shows the accuracy and classification report of evaluating the ALexNet model by using the same method we did for MobileNetV2 with a held out Jaffe validation split (42 samples, 20% of the dataset). From the table, it can be observed that the accuracy is around 55%. Which shows genuine difficulty of generalizing to JAFFE given its small size of 213 images and domain shift from FER2013. The result is more in align with [1].

From observing the values of each class, it is observed that the strongest class was surprise, which achieved an F1 score of 0.89 with a perfect recall of 1.00, correctly identifying all the 8 samples in the validation set. Fear on the other hand achieved a perfect precision of 1.00, meaning every prediction of fear was correct. However, its recall of 0.50 indicates that the model missed half the actual fear samples. Similarly sad showed a high precision of 1.00 but a low recall of 0.30, suggesting the model was overly conservative in predicting the sad expression. The weakest performing class were disgust with an F1 score of 0.22 and angry with an F1 score of 0.31, both of which had few support samples, making reliable classification difficult.

Fig 29 shows the confusion matrix for the model. The evaluation showed model shows strong bias towards predicting the neutral and surprise. Angry was correctly predicted only 2 out of 5 times, with three samples misclassified as neutral. Sad was the most challenging class overall, with only 3 out of 10 samples correctly identified. Disgust was correctly predicted only 1 out of 4 times, with the remaining three misclassified as angry.

These results highlight the limitations of applying a large model like AlexNet to a very small and domain specific dataset like JAFFE. The drop in performance compared to the previously reported

figure is attributed to the correction of the evaluation methodology, where the model is now assessed strictly on the held out validation split rather than the full dataset, which eliminates the data leakage.

3.7. Subnetwork Observation and Results

3.7.1. Architecture Search Result

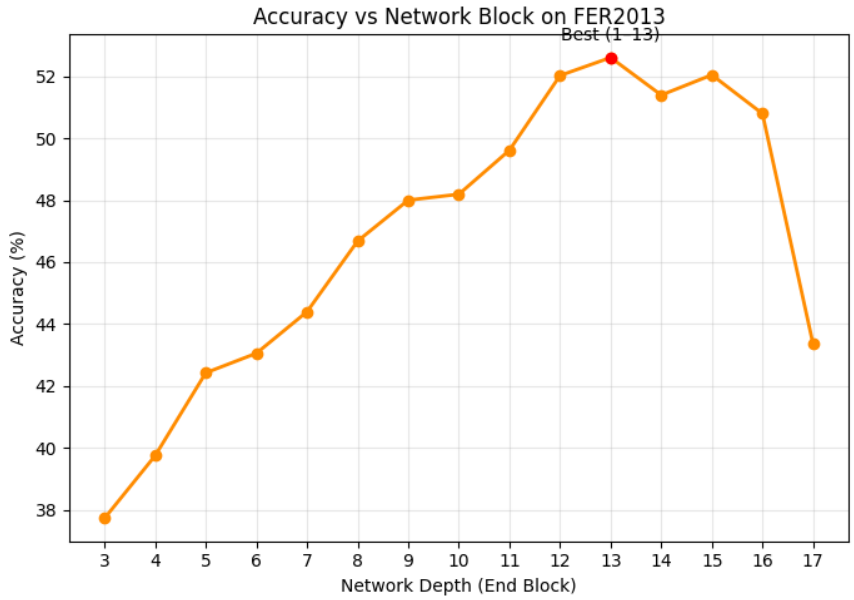

Figure 30. Accuracy vs the block depth in FER2013

Evaluation and Result

An exhaustive search was conducted on the FER2013 dataset to identify the most optimal subnetwork from MobileNetV2. All candidate subnetwork from blocks 1-3 up to block 1-17 were evaluated. The accuracy steadily improved from 37.98% at blocks 1-3 and peaked at blocks 1-13, after which the performance began to decline.
The best performing subnetwork consists of blocks 1-13, achieving a mean linear probing accuracy of 52.60%, while reducing the parameter count from 2.23M to 0.99M, which is a 55.8% reduction compared to the full model.

This indicates that the intermediate layer captures the most descriptive features for facial expression recognition, while the deeper layers may introduce redundancy that does not generalize well under limited data conditions.

3.7.2. Search Stability Analysis

Table 8. Top five performing subnetwork

Blocks	Mean %	Std %	Wins	Seed 42	Seed 123	Seed 456	Seed 789	Seed 999
1-13	53.11	0.58	4/5	52.65	53.44	53.30	52.28	53.90
1-12	51.94	0.30	1/5	51.76	52.00	51.49	52.37	52.09
1–15	51.69	0.44	0/5	51.58	52.07	51.21	52.32	51.28
1-14	51.50	0.41	0/5	50.86	51.32	52.07	51.51	51.74
1-16	50.78	0.54	0/5	50.74	50.26	51.11	51.63	50.19

Evaluation and Result

To evaluate the robustness of the search results, the experiment was repeated across five random seeds using a train/validation split, where the test set was kept completely unseen during the search phase.

Two non-overlapping 15% subsets were drawn from the training data, one for validation and the other for the logistic regression classifier. From table 8, it can be observed that, the subnetwork consisting of blocks 1-13 achieved the highest accuracy in four out of five seeds, with accuracies of 52.65%, 53.44%, 53.30%, 52.28% and 53.90% respectively. The path achieved a mean accuracy of 53.11% and the standard deviation of 0.58% indicates the low variance and better stability. Path 1-12 was able to achieve an accuracy of 52.37% in seed 789, when compared to the path 1-13 is only marginal difference of 0.09% which is in favour of block 1-12 over block 1-13.

This shows that the block 1-13 is more reliable choice overall. But the deeper blocks such as blocks 1-15 showed slightly lower mean performance despite having higher parameter count, which further supports the effectiveness of the proposed subnetwork selection strategy.

3.7.3. Limited Data Evaluation on FER2013

Table 9. Data fraction Comparison table of base and efficient model

Fraction	N	Baseline	Evolved	Difference	p
5%	1435	49.27%±0.53%	46.38%±1.66%	-2.89%±1.54%	0.0198
10%	2870	52.84%±0.56%	52.32%±0.29%	-0.52%±0.65%	0.1857
20%	5741	57.55%±0.33%	57.14%±1.18%	-0.41%±1.37%	0.5810
50%	14354	62.73%±0.41%	61.92%±0.15%	-0.81%±0.54%	0.0393
100%	28709	67.00%±0.24%	65.27%±0.27%	-1.72%±0.43%	0.0013

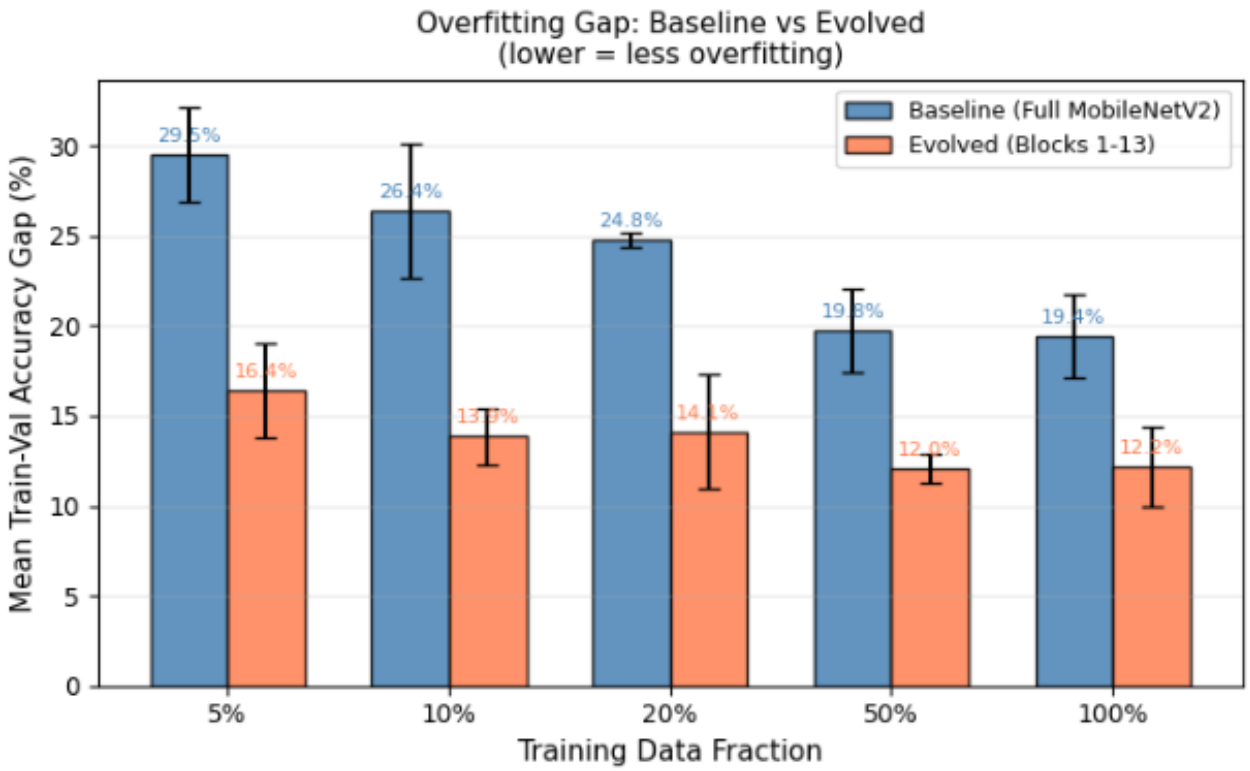

Figure 31. Train - validation accuracy gap between baseline and evolved model on FER2013 data fractions

Evaluation and Result

Table 9. represents the performance comparison of both the full MobileNetV2 baseline model and the proposed efficient subnetwork model under limited the limited data across five fractions.

At 10% and 20% of the data fractions, no significant difference was observed between the baseline and the proposed subnetwork ($p = 0.1857$ and $p = 0.5810$ respectively), indicating the two models are not statistically significant at these scales despite being 55.8% parameter reduction. At 5%,50% and 100% data fractions, a statistically significant accuracy gap can be observed between the two models, which is expected as the full model has more parameters. Despite this, the evolved subnetwork remains competitive across all fractions while maintaining lower overfitting behaviour.

Across all the data fractions, the evolved subnetwork consistently exhibits a lower train validation accuracy gap compared to the full baseline. At 5% data, the baseline exhibited a train-validation gap of 29.48% compared to 16.43% for the subnetwork model, which is a difference of 13.05%. At 10% data fraction the gaps were 26.42% and 13.85% respectively, a difference of 12.57 percentage points. At 20% data, the gaps are 24.79% and 14.14% respectively which is a difference of 10.65%. At 50% data, the gaps were 19.76% and 12.04%, which has a difference of 7.72% and finally at 100%, the gaps were 19.42% and 12.17% respectively, which is a difference of 7.25%. This consistent reduction in train-validation gap across the different fractions indicates that the compact subnetwork overfits substantially less than the full model regardless of the data scale, suggesting that the reduction of parameter count act as a form of regularization.

Overall, the proposed subnetwork was able to demonstrate a strong parameter reduction, competitive accuracy in a moderate low data setting and was able to consistently reduce overfitting across all evaluated conditions making it a suitable choice for resource constrained deployment.

3.7.4. Subject Independent Evaluation on CK+

Table 10. Classification table for evolved subnetwork on CK+ dataset

Emotion	Precision	Recall	F1- score	Support
Anger	0.8491	0.7778	0.8119	405
Contempt	0.8154	0.6543	0.7260	162
Disgust	0.9224	0.9623	0.9419	531
Fear	0.8186	0.8622	0.8398	225
Happy	0.9138	0.9903	0.9505	621
Sadness	0.8025	0.7579	0.7796	252
Surprise	0.9851	0.9759	0.9805	747
Accuracy			0.9042	2943
Macro Avg	0.8724	0.8544	0.8615	2943
Weighted Avg	0.9023	0.9042	0.9020	2943

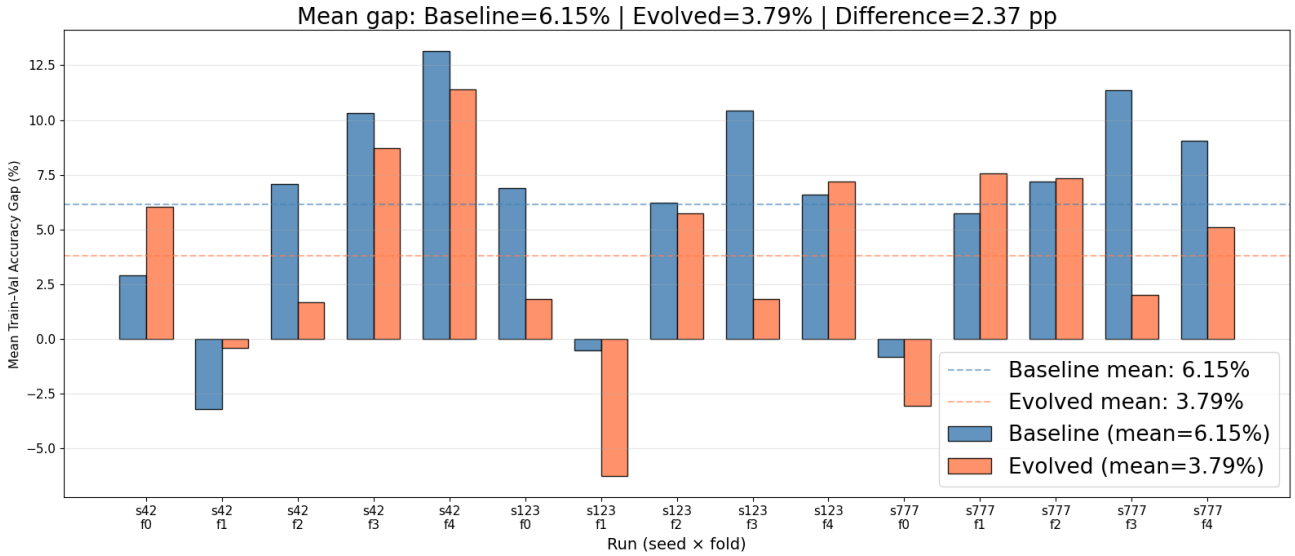

Figure 32. Mean train-val accuracy gap comparison between the full baseline model and subnetwork across 15 evaluation runs on CK+(3 seeds x 5 folds). Lower value indicates less overfitting

Evaluation and Result

The efficient subnetwork model was able to achieve an overall accuracy of 90.47% compared to the baseline of 88.58%, which is a difference of +1.89%. While this difference did not reach statistical

significances as t = 1.607 and p = 0.130; Wilcoxon W = 23.0, p = 0.116, the per class analysis shows improvement across most of the emotion categories.

The evolved model was able to improve six out of the seven emotion categories by F1 score. The most notable gains were observed in contempt, where F1 improved from 0.659 to 0.726 and anger where F1 improved from 0.784 to 0.812. The macro F1 improved from 0.834 to 0.862, which indicates broader class level improvement across most of the categories. Fear on the other hand showed a marginal improvement from 0.837 to 0.840 , while sadness improved from 0.727 to 0.780. Happy showed a small decrease in precision but maintained near perfect recall and surprise remained the strongest class for both models.

From the classification report table 10, it can be observed that the evolved model was able to achieve a recall of 0.990 for happy class, meaning it correctly identified the vast majority of happy samples from all the 15 evaluation runs. This is likely due to happy being the most visually distinct emotion in CK+ dataset, identified by a clear upward movement of the lip corner and cheek raising, which makes it easier for both the models to recognize it consistently. The baseline also achieved a very high recall of 0.992 for happy, which confirms that this emotion is reliably detected regardless of the model size. The high performance on happy and surprise, which achieved an F1 score of 0.951 and 0.981 for the evolved model, is consistent with the findings from the other baseline experiments in this study, where these two emotions were performing strongly across all methods and datasets.

An analysis of the train-val accuracy gap across all the 15 evaluation runs also further supports the generalization advantage of the evolved subnetwork. This can be seen from the Fig 31, the baseline showed a mean train-val gap of 6.15% when compared to the 3.79% for the evolved subnetwork model, which is a difference of 2.37%. The evolved model was able to exhibit a lower overfitting in 11 out of the 15 runs as shown from the figure. A small number of runs exhibited negative train-val gaps, which occurs when the validation fold contains subject whose expressions are more easily classified than those in the training partition, which is an expected artifacts of subject level splitting on a small dataset.

The variance which can be observed across the folds is partly because of the small size of the CK+, which contains only 981 images across 118 subjects, with an average of approximately 8 images per subject. Under subject independent splitting, the individual folds can vary in class representation because for minority classes such as contempt will have only 54 images in total. In such cases the full baseline model is more prone to memorizing subject specific patters because of the higher parameter count, while the smaller evolved subnetwork is less prone to this behaviour.

Overall, it can be said that the evolved subnetwork was able to achieve a competitive accuracy on CK+ with 55.8% fewer parameter and demonstrated lower overfitting across the majority of the emotion categories, which is consistent with the findings on Fer2013. The absence of statistical significance is expected as the dataset size is small and high fold level variance inherent to subject independent evaluation on CK+, and does not diminish the practical value of parameter reduction

3.8. Cross Method Comparison

Table 11. Represents different methods along with proposed method

Method	Dataset	Accuracy	Parameters	Strategy
Two-Stage Transfer Learning [1]	FER2013 & EmotiW	55.6%	High	Two Stage Fine tuning.
MobileNetV2 Transfer Learning [2]	JAFFE	85.54%	Full Model	Single stage fine tuning.
Improved MobileNetV2 [3]	FER2013	70.21%	Low	Lightweight + Attention Mechanism
Improved MobileNet V2 [3]	CK+	98.12%	Low	Lightweight + Attention Mechanism
PathNet [12]	SAVEE/EmoDB	94%	Medium	Subnetwork + GA
MobileNetV2Baseline (this paper)	JAFFE	77.27%	Full	Full fine - tuning
ResNet50 (this Paper)	FER2013	66%	High	Full fine tuning
Two Stage MobileNet V2 (this work)	JAFFE + FER2013	70.43%	Full Model	Two stage fine tuning
Two Stage AlexNet(this work)	JAFFE	55%	High	Two stage fine tuning
Proposed Subnetwork	FER2013	65.27%	0.99M	Subnetwork Selection
Proposed Subnetwork	CK+	90.47%	0.99M	Subnetwork Selection

Evaluation and Result

The Table 11 above represents the comparison of the different methods that have been used. From this table several observations can be made. First, direct comparison accuracy across all method is limited as different methods were evaluated on different datasets. The improved MobileNetV2 [3], achieved the highest accuracy on CK+ at 98.12%, compared to the 90.47% for the proposed subnetwork. However, this is achieved by adding attention mechanism to the full mode, which further increases the architectural complexity. The proposed subnetwork takes the opposite approach by

removing the later blocks to reduce the complexity, which is more in alignment with the goal, which is parameter efficient deployment.

Second, among the methods evaluated in the study, the two stage AlexNet achieved an accuracy of 58% on JAFFE under subject independent evaluation, which is comparable to the 55.6% reported by [1] using a similar two stage approach on FER2013. JAFFE contains only 213 images across 10 Japanese female subjects making evaluation results highly sensitive to random subject alignment and less representative of the real-world deployment conditions. The accuracy on JAFFE should therefore be interpreted with caution as it may not reflect performance on more diverse and challenging dataset. In contrast, FER2013 with 28,709 images and CK+ with 981 images across 118 subjects provide more reliable benchmarks for evaluating generalization.

Third, the most important contribution of the proposed subnetwork is not raw accuracy, but the combination of parameter efficiency and reduced overfitting. The proposed subnetwork achieves 55.8% fewer parameters than the full MobileNetV2 baseline while producing no statistically significant difference at 10-20% data fractions on FER2013 and achieving competitive accuracy of 90.47% on CK+. Furthermore, the train-val accuracy gap of the proposed subnetwork is consistently lower than the full baseline across all evaluated conditions, which confirms that the compact architecture generalizes more reliably under limited data.

Fourth, the PathNet based approach achieves a 94% on SAVEE and EmoDB but relies on genetic algorithm optimization which is computationally expensive. The proposed method performs an exhaustive grid search using linear probing, requiring no gradient based optimization during the search phase. This makes it significantly more practical for resource constrained settings where computational efficiency is important.

Finally, among the transfer learning approaches evaluated in this study, the two stage methods produced results consistent with published literature when evaluated under subject independent conditions, confirming that sequential domain adaptation is a valid strategy for small target datasets. However, all transfer learning techniques require full model training and do not address parameter efficiency. The proposed subnetwork selection strategy addresses this gap directly by identifying a subnetwork that achieves competitive accuracy while substantially reducing the model complexity by 55.8%.

3.9. Final Observation

From the experiments conducted in this research, several key findings emerged across all methods and datasets.

Overfitting was consistently observed across all methods, particularly in larger models such as ResNet50 and the full MobileNetv2 when trained on limited data. The train-validation gap analysis confirms that larger models tend to memorize training data rather than learning generalizable features, which is reflected in the consistently higher train-val gaps observed for the baseline model across all data fractions on FER2013 and across all folds on CK+.

All models consistently struggled with visually similar expressions such as anger, disgust and sadness. These three emotions share subtle facial muscle movements especially around the brow, nose and mouth regions that are difficult to distinguish under limited data conditions. In contrast strong performance was observed for happy and surprise across all methods and datasets, which aligns with the established understanding that these emotions have more pronounced and visually distinct facial movement. This finding was consistent across FER2013, JAFFE and CK+, suggesting that it is a genuine property of these expression classes rather than dataset specific artifacts.

The proposed subnetwork selection strategy demonstrated that a carefully chosen compact subnetwork can achieve competitive accuracy while substantially reducing overfitting compared to the full model. The train-val gap analysis confirmed that the evolved subnetwork consistently overfits less than the full MobileNetV2 baseline across all evaluated data fractions on Fer2013 and across all folds on CK+. This is particularly valuable in real world deployment scenarios where training data is scarce and model size matters.

The stability analysis of the search procedure confirmed that the proposed approach is reliable and reproducible. The optimal subnetwork consisting of blocks 1-13 was identified consistently across four out of five random seeds on FER 2013, with the single block which had a disagreement with a margin of 0.09%. The additional finding that JAFFE is too small for reliable search while FER2013 is not provides a practical guideline for applying this method to other datasets.

Overall, the results confirm that optimal performance does not require the most complex model. A carefully selected subnetwork can match or approach full model accuracy while offering substantial advantages in parameter efficiency, overfitting resistance and deployment suitability for resource constrained environments.

3.10. Limitations

Despite the promising results demonstrated in this study, several limitations should be acknowledged.

First, the search space was restricted to prefix subnetworks, meaning only consecutive blocks starting from block 1 through k, where $k \in \{3 \ldots\ldots 17\}$,meaning the path always starts at block 1 and only the depth is varied. More flexible search strategies allowing non-consecutive block selection could potentially identify more efficient configurations.

Second, the proposed subnetwork was only evaluated on MobileNetV2. Whether the optimal truncation generalizes to other architecture such as MobileNetV3 or EfficientNet remains an open question.

Third, the JAFFE dataset contains only 213 images of the Japanese female subjects, limiting the generalizability of the JAFFE results to broader populations and real-world conditions.

The subnetwork search on CK+ was conducted using a single seed with a subject independent subset split. While the search independently confirmed the blocks 1-13 as optimal on CK+, the search stability across multiple seed was not evaluated on CK+ as it was on FER2013. Given the small size of the CK+, search results on the dataset may be less reliable than those obtained on FER2013.

Finally, all experiments were conducted on a single consumer grade GPU which limited the scale of experiments, particularly the number of seeds and data fractions evaluated.

Conclusions

1. This research investigated facial expression recognition under limited data conditions, evaluating multiple transfer learning approaches and proposing a parameter efficient subnetwork selection strategy based on exhaustive search within a pretrained MobileNetV2 backbone.
2. From the literature analysis, multiple expression recognition techniques were identified and analysed, providing the foundation for selecting appropriate methods, preprocessing strategies and augmentation techniques for the experimental phase. The analysis highlighted that while deep learning approaches have achieved stronger results on large scale datasets, performance degrades significantly under limited data conditions due to overfitting.
3. Four transfer learning approaches were implemented and evaluated which were single stage MobileNetV2 fine-tuning, ResNet50 transfer learning, two stage MobileNetV2 fine tuning and AlexNet finetuning. Each method demonstrated different trade-offs between accuracy, dataset requirements and overfitting behaviour. The two stage approaches produced results consistent with the published literature when evaluated under subject independent conditions, confirming that sequential domain adaptation is a valid strategy for small datasets.
4. A structured subnetwork selection strategy was proposed and evaluated, identifying blocks 1-13 of MobileNetV2 as the optimal subnetwork for FER under limited data conditions. The selected subnetwork achieved a 55.8% parameter reduction while producing no statistically significant accuracy difference at 10-20% data fractions on FER2013 and achieving competitive accuracy of 90.47% on CK+.
5. The train-val gap analysis confirmed that the evolved subnetwork overfits substantially less than the full model across all the evaluated data fractions, with a gap reduction ranging from 7.25 to 13.05% on FER2013 and a mean reduction of 2.37% on CK+. This confirms that model compactness provides regularization benefits which are valuable under limited data conditions.
6. The stability analysis confirmed that the search procedure reliably identifies blocks 1-13 across four out of five random seeds. This provides confidence in the reproducibility of the proposed approach when applied to datasets of similar or larger scale.

References

1. Ng, Hong-Wei and Nguyen, Viet Dung and Vonikakis, Vassilios and Winkler, Stefan. Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning*In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*2015. Available from: https://dl.acm.org/doi/10.1145/2818346.2830593..

2. Barile, Paolo, Clara Bassano, and Paolo Piciocchi. Transfer Learning for Facial Emotion Recognition on Small Datasets. *Journal of Systemics, Cybernetics and Informatics* (2024), pp. 1–5. Available from: https://www.researchgate.net/publication/381601170_Transfer_Learning_for_Facial_Emotion_Recognition_on_Small_Datasets..

3. Yali and Zhang YU Chengxun. Facial Expression Recognition Based on Improved MobileNetV2 Network Model*In: 2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT)*2024. Available from: https://ieeexplore.ieee.org/document/10962869..

4. ROY, Shuvendu and Etemad, Ali. Active Learning with Contrastive Pre-training for Facial Expression Recognition (2023). Available from: https://arxiv.org/abs/2307.02744..

5. ISTIQOMAH, Annisa; SARI, Atika; SUSANTO, Ajib and RACHMAWANTO, Eko. Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning Resnet-50. *Journal of Applied Informatics and Computing*, vol. 8 (2024), pp. 257–264. Available from: https://www.researchgate.net/publication/386590220_Facial_Expression_Recognition_using_Convolutional_Neural_Networks_with_Transfer_Learning_Resnet-50. [viewed May 20, 2026].

6. LI, Bin and LIMA, Dimas. Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, vol. 2 (2021), pp. 57–64. Available from: https://www.sciencedirect.com/science/article/pii/S2666307421000073. [viewed May 21, 2026].

7. REVINA, I. Michael and EMMANUEL, W. R. Sam. A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University - Computer and Information Sciences*, vol. 33 (2021), no. 6, pp. 619–628. Available from: https://www.sciencedirect.com/science/article/pii/S1319157818303379. [viewed May 21, 2026].

8. GE, Huilin; ZHU, Zhiyu; DAI, Yuewei; WANG, Biao and WU, Xuedong. Facial expression recognition based on deep learning. *Computer Methods and Programs in Biomedicine*, vol. 215 (2022), pp. 106621. Available from: https://www.sciencedirect.com/science/article/pii/S0169260722000062. [viewed May 21, 2026].

9. Yating YU; Yiliu Sun and Zejia Yang. An Unsupervised Facial Expression Recognition Method Based on CycleGAN, pp. 669–674.January 1, 2022. Available from: https://ieeexplore.ieee.org/document/9758435. [viewed May 21, 2026].

10. ZHUANG, Fuzhen; QI, Zhiyuan; DUAN, Keyu; XI, Dongbo; ZHU, Yongchun, et al. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, vol. 109 (2021), no. 1, pp. 43–76. Available from: https://ieeexplore.ieee.org/document/9134370. [viewed May 21, 2026].

11. HOSPEDALES, Timothy; ANTONIOU, Antreas; MICAELLI, Paul and STORKEY, Amos. *Meta-Learning in Neural Networks: A Survey.* [2020November 7]. Available from: http://arxiv.org/abs/2004.05439. [viewed May 21, 2026].

12. NGUYEN, Dung; SRIDHARAN, Sridha; NGUYEN, Duc T.; DENMAN, Simon; DEAN, David, et al. *Meta Transfer Learning for Emotion Recognition.* [2020June 23]. Available from: http://arxiv.org/abs/2006.13211. [viewed May 21, 2026].

13. DENG, Shuchao; SUN, Yanan and GALVAN, Edgar. *Neural Architecture Search using Genetic Algorithm for Facial Expression Recognition.* [2023April 12]. Available from: http://arxiv.org/abs/2304.12194. [viewed May 21, 2026].

14. RAJESWARAN, Aravind; FINN, Chelsea; KAKADE, Sham and LEVINE, Sergey. *Meta-Learning with Implicit Gradients.* [2019September 10]. Available from: http://arxiv.org/abs/1909.04630. [viewed May 21, 2026].

15. PENG, Sisi and ZHENG, Kaining. Application of Meta Learning in Face Recognition. *Journal of Physics: Conference Series*, vol. 1757 (2021), pp. 012036. Available from: https://www.researchgate.net/publication/349018222_Application_of_Meta_Learning_in_Face_Recognition/citation/download. [viewed May 21, 2026].

16. ROY, Shuvendu and ETEMAD, Ali. *Analysis of Semi-Supervised Methods for Facial Expression Recognition.* [2022July 31]. Available from: http://arxiv.org/abs/2208.00544. [viewed May 21, 2026].

17. HINTON, Geoffrey; KRIZHEVSKY, Alex and SUTSKEVER, Ilya. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105. Available from: https://www.researchgate.net/publication/319770183_Imagenet_classification_with_deep_convolutional_neural_networks. [viewed May 21, 2026].

18. HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing and SUN, Jian. *Deep Residual Learning for Image Recognition.* [2015December 10]. Available from: http://arxiv.org/abs/1512.03385. [viewed May 21, 2026].

19. ALI, Luqman; ALNAJJAR, Fady; JASSMI, Hamad Al; GOCHO, Munkhjargal; KHAN, Wasif, et al. Performance Evaluation of Deep CNN-Based Crack Detection and Localization Techniques for Concrete Structures. *Sensors*, vol. 21 (2021), no. 5, pp. 1688. Available from: https://www.mdpi.com/1424-8220/21/5/1688. [viewed May 21, 2026].

20. Yexiu ZHONG; Senhui Qiu; Xiaoshu Luo; Zhiming Meng and Junxiu Liu. Facial Expression Recognition Based on Optimized ResNet, pp. 84–91.2020-06-01. Available from: https://www.researchgate.net/publication/343026416_Facial_Expression_Recognition_Based_on_Optimized_ResNet. [viewed May 21, 2026].

21. Shan LI; Weihong Deng and JunPing Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild, pp. 2584–2593.July 1, 2017. Available from: https://ieeexplore.ieee.org/document/8099760. [viewed May 21, 2026].

22. MAHARANA, Kiran; MONDAL, Surajit and NEMADE, Bhushankumar. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, vol. 3 (2022), no. 1, pp. 91–99. Available from: https://www.sciencedirect.com/science/article/pii/S2666285X22000565. [viewed May 21, 2026].

23. ZHAO, Zehui; ALZUBAIDI, Laith; ZHANG, Jinglan; DUAN, Ye and GU, Yuantong. A comparison review of transfer learning and self-supervised learning: Definitions, applications,

advantages and limitations. *Expert Systems with Applications*, vol. 242 (2024), pp. 122807. Available from: https://www.sciencedirect.com/science/article/pii/S0957417423033092. [viewed May 21, 2026].

24. XU, Mingle; YOON, Sook; FUENTES, Alvaro and PARK, Dong Sun. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*, vol. 137 (2023), pp. 109347. Available from: https://www.sciencedirect.com/science/article/pii/S0031320323000481. [viewed May 21, 2026].

25. SANDLER, Mark; HOWARD, Andrew; ZHU, Menglong; ZHMOGINOV, Andrey and CHEN, Liang-Chieh. *MobileNetV2: Inverted Residuals and Linear Bottlenecks.* [2019March 21]. Available from: http://arxiv.org/abs/1801.04381. [viewed May 21, 2026].

26. ZHANG, Wancong; GX-CHEN, Anthony; SOBAL, Vlad; LECUN, Yann and CARION, Nicolas. *Light-Weight Probing of Unsupervised Representations for Reinforcement Learning.* [2024May 31]. Available from: http://arxiv.org/abs/2208.12345. [viewed May 21, 2026].

27. M. LYONS; S. Akamatsu; M. Kamachi and J. Gyoba. Coding Facial Expressions with Gabor Wavelets, pp. 200–205.April 1, 1998. Available from: https://ieeexplore.ieee.org/document/670949. [viewed May 21, 2026].

28. LYONS, Michael J. *"Excavating AI" Re-Excavated: Debunking a Fallacious Account of the JAFFE Dataset.* [2021July 28]. Available from: http://arxiv.org/abs/2107.13998. [viewed May 21, 2026].

29. THARWAT, Alaa. Classification Assessment Methods: a detailed tutorial (2018). Available from: https://www.researchgate.net/publication/327148996_Classification_Assessment_Methods_a_detailed_tutorial. [viewed May 21, 2026].

Annexes

1.1 Annex1. Accepted Conference Paper

The following paper was accepted for presentation at the International Conference on Information and Software Technologies (ICIST 2026),15-16 October 2026, Kaunas, Lithuania.

Acceptance confirmation email received 12 May 2026:

From: icist2026-0@easychair.org <icist2026-0@easychair.org> on behalf of ICIST 2026 <icist2026-0@easychair.org>
Sent: Tuesday, May 12, 2026 8:43 PM
To: Ananthakrishnan Thuruthel Murali <ananthakrishnan.thuruthel@ktu.edu>
Subject: ICIST 2026 notification for paper 7923

Dear Ananthakrishnan Thuruthel Murali,

We are happy to let You know that Your paper No.7923 "Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low-Data Regimes" has been accepted for presentation at the International Conference on Information and Software Technologies (ICIST 2026) Conference, which will be held on 15-16 October in Kaunas, Lithuania.

Your submission received favourable feedback from the reviewers and was recommended for acceptance for its clear and well-organized approach. We sincerely appreciate the time and effort You invested in preparing your work and congratulate You on this achievement. We look forward to Your presentation at the Conference and believe that Your contribution will encourage meaningful discussions and support further developments in the field.

Please note that information regarding final paper submission, guidelines for presenters, and payment instructions will be shared next week by a separate email.

If You have any questions or concerns, please do not hesitate to contact us by email at icist@ktu.lt.

Regards,
Programme Committee
icist@ktu.lt

SUBMISSION: 7923
TITLE: Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low-Data Regimes

---------- REVIEW 1 ----------

SUBMISSION: 7923
TITLE: Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low-Data Regimes

---------- Overall evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
The paper introduces a practical idea of selecting a compact subnetwork within MobileNetV2 for facial expression recognition under limited data conditions.
A strong aspect is the clear experimental setup with multiple datasets (FER2013, CK+) and the use of subject-independent validation, which improves reliability. The consistent parameter reduction (~56%) together with accuracy gains is also a meaningful contribution for resource-constrained scenarios.
The method is relatively simple (truncating first k blocks), and the novelty compared to existing pruning or NAS approaches is limited. The search strategy evaluates subnetworks using only 15% of the data without augmentation, potentially biasing selection and failing to reflect the final training conditions. The reported improvements on FER2013 are small (often below 1–1.5%), raising questions about practical significance despite statistical tests. The paper lacks deeper analysis of why blocks 1–13 are optimal (no feature-level or representation analysis).
The work is well-structured and practically motivated, but methodological novelty and analytical depth remain moderate.

---------- REVIEW 2 ----------

SUBMISSION: 7923
TITLE: Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low-Data Regimes

---------- Overall evaluation ----------
SCORE: 2 (accept)
----- TEXT:
This paper presents a well-structured and practically relevant approach to parameter-efficient facial expression recognition under low-data conditions. The proposed deterministic subnetwork selection strategy within MobileNetV2 is both elegant and computationally efficient, avoiding the complexity of traditional neural architecture search methods while still achieving substantial parameter reduction. I particularly appreciated the thorough experimental design, including stability analysis across multiple random seeds, evaluation under varying data fractions, and subject-independent validation on CK+. The results convincingly demonstrate that the selected compact subnetwork maintains competitive accuracy while significantly reducing overfitting and model complexity, which is highly valuable for deployment in resource-constrained environments. Overall, the work provides a meaningful contribution to low-data FER research and highlights an effective direction for efficient deep learning model design.

Parameter-Efficient MobileNetV2 Subnetwork Selection for Facial Expression Recognition in Low-Data Regimes

Ananthakrishnan Thuruthel Murali [0009-0004-0425-7839] and Armantas Ostreika

[0000-0001-5718-3766]

[1] Kaunas University of Technology, Faculty of Informatics; anathu@ktu.lt (A.T.M.)
[2] Kaunas University of Technology, Faculty of Informatics, Department of Multimedia; armantas.ostreika@ktu.lt(A.O.)

Abstract. Facial Expression Recognition (FER) in low data regimes is challenging, as deep models with larger parameter counts tend to overfit when trained on limited data. While full model fine-tuning is effective at scale, it is often suboptimal in such scenarios due to inefficient parameter utilization. This work proposes a structured prefix subnetwork selection strategy within a pretrained MobileNetV2, where candidate subnetworks are evaluated efficiently via frozen feature extraction and logistic regression, without any gradient based optimization during search. The approach is specifically designed for parameter efficient FER under limited data conditions. The candidate subnetworks are constructed by selecting the first k consecutive inverted residual blocks ($k \in \{3, \ldots \ldots, 17\}$) and evaluated on a fixed 15% subset using a frozen backbone and logistic regression classifier. An exhaustive search on FER2013 identified blocks 1-13 as the most optimal blocks and achieved 55.8% reduction in parameters (2.23Mto 0.99M) while achieving statistically equivalent accuracy to the full MobileNetV2 baseline at 10-20% data fractions with substantially lower overfitting. Robustness is confirmed across five random seeds, with the blocks 1-13 selected in four of five seeds. Furthermore, Subject independent 5-fold cross validation on CK+ yields a +1.89% mean accuracy difference over the baseline, while improving macro level performance across 6 of 7 emotion categories. These results demonstrate that a carefully selected subnetwork can achieve competitive generalization with significantly improved parameter efficiency compared to full model fine-tuning, making it well suited for resource constrained and data scarce deployment.

Keywords: Facial Expression Recognition, Low-Data Learning, Parameter-Efficient Learning, Subnetwork Selection, MobileNetV2, Linear Probing.

1 Introduction

Facial Expression Recognition (FER) is an important area within computer vision [1], as it enables different applications such as emotion detection, virtual reality interactions, user experience optimization and security systems. An accurate FER system can help improve the human – computer interaction by making technology more adaptive and responsive. FER datasets typically include seven basic emotion categories namely happy, sad, fearful, anger, surprised, disgust and neutral [7]. A standard FER pipeline consists of three main stages: preprocessing, feature extraction and classification [8]. In FER pipelines involving deep learning models, preprocessing is applied to raw images prior to feature extraction , while feature extraction and classification are typically performed jointly by the model [4].

While there is an increase in the number of deep learning techniques, which has made significant advancement in the field, many of these approaches rely on large scale annotated datasets to achieve high performance [1]. For instance, state of the art results on FER2013 datasets have been achieved using deep convolutional architectures trained with extensive data and fine-tuning strategies [5]. However, this becomes a challenge in many real-world scenarios where such large datasets are not available primarily due to privacy constraints, annotation costs and domain specific limitations. As a result, FER systems frequently operate in low data constraints where the deep models tend to overfit and exhibit poor generalization. Addressing FER under limited data conditions is therefore a critical and underexplored problem. Existing approaches to mitigate this issue include transfer learning [11] [12] [13],data augmentation [14] and use of lightweight models. Few shot learning (FSL) has also

emerged as a promising approach to address these limitations by letting models learn from a small number of samples while maintaining generalization performance [10].

The recent research has explored the idea of learning a task specific pathway within a network, where only a subset of parameters are activated for a given task, an example would be [16] the pathnet based transfer learning approach which learns the optimal pathways through neural networks to improve the knowledge transfer across domains. While effective, these approaches rely on complex optimization strategies, such as genetic algorithms, which are computationally expensive.

To enable a comprehensive evaluation under different conditions, experiments in this work are conducted on two different datasets, FER2013 and CK+. FER2013 contains facial images collected in unconstrained environments, making it suitable for robustness in real world scenarios [12] [13], whereas the CK+ is smaller, lab controlled dataset used for benchmarking under subject independent evaluation protocol [14] [15]. Using both datasets allows us to assess the proposed approach in both realistic and controlled settings, particularly under low data conditions.

This paper focuses on evaluating and improving an FER performance under limited data by leveraging transfer learning, data augmentation and lightweight convolution neural networks. However, these strategies typically involve retraining or fine-tuning entire pretrained models, which can be suboptimal in low-data settings due to overfitting and inefficient parameter usage [16].

In this work, we have proposed a structured, subnetwork selection strategy that identifies a compact and parameter-efficient subnetwork within a pretrained MobileNetV2 for FER under limited data constraints. Rather than fine-tuning a full network, we formulate the problem as a structured search over the subnetworks, each formed by retaining only the first k consecutive inverted residual blocks. Candidates subnetworks are evaluated efficiently via linear probing by extracting the features once from the frozen backbone and training a logistic regression classifier which avoids the computational cost of training a full model during search. The key contribution of this work are as follows: (1) a structured subnetwork selection strategy for low data fer, where candidate subnetworks are formed by retaining the first k inverted residual blocks of a pretrained MobileNetV2 and evaluated efficiently via frozen feature extraction and logistic regression, avoiding full end to end training during search; (2) empirical demonstration that the selected subnetwork achieves 55.8% parameter reduction while remaining statistically equivalent to the full model at 10-20% data fractions; (3) evidence of strong search stability across five random seeds with blocks 1-13 selected in four of five runs.

As such the proposed method focuses on a structured and efficient subnetwork selection strategy at block level instead of the neuron level, for low data FER.

2 Related Works

Transfer learning has been widely adopted as a primary strategy for Fer under limited conditions, typically by fine tuning large pretrained models on target datasets. For example, the use of AlexNet and VGG-CNN-M-2048 models, [17] pretrained on ImageNet dataset to perform a two-stage finetuning, first on the FER 2013 dataset, followed by a second stage fine tuning on the EmotiW dataset, achieving 55.6% accuracy compared to a 39.13 baseline. Similarly, in [11] the authors employ a general transfer learning using a MobileNetV2 model by replacing the final classification

layer to match the seven emotion classes of the JAFFE dataset. To mitigate the issue of overfitting, the dataset is expanded to around 79,450 using data augmentation and were able to achieve an accuracy of 85.54% by fine tuning the entire model. While effective, both approaches adapt the full model and depend on large, augmented datasets, which limit their applicability in genuinely data scarce settings.

A parallel line of research focuses on designing lightweight CNN architecture that reduces computational cost while maintaining competitive FER accuracy. For instance, [15] proposes an enhanced version of MobileNetV2, by incorporating inverted residual structures, linear bottlenecks and attention mechanisms to improve the feature representation while maintaining low computational complexity, while [18] presents an optimized ResNet variant for FER. However, these approaches still require full model training and do not specifically address parameter efficiency under limited data conditions.

Another line of research explores task specific subnetworks selection and automated architecture search. In [16] the author proposes learning the optimal pathway through a neural network, using evolutionary strategies, such as genetic algorithms, activating only subsets of layers per task. This concept is closely related to Neural Architecture Search (NAS), [23] which aims to automatically identify the efficient network structure. However, both PathNet and NAS methods typically involve computationally expensive and complex search procedures, limiting their applicability in low-data and resource constrained settings. In contrast the proposed method performs a deterministic prefix search via frozen feature extraction and logistic regression, requiring no gradient based optimization during search, making it more suitable for low data and resource constrained FER.

Table 12. Comparison of FER Methods in terms of Accuracy and Model Complexity

Method	Model	Datasets	Accuracy(%)	Parameters	Strategy
[17]	AlexNet/VGG-CNN-M	FER2013 & EmotiW	55.6	High	TransferLearnign(2Stage)
[11]	MobileNetV2	JAFFE	85.54	Low	TransferLearning+ Augmentation
[15]	Improved MoblenetV2	FER2013 & CK+	70.21 & 98.12	Low	Lightweight+Attention
[16]	Custom PathNet	SAVEE/Emo-DB	94%	Medium	Subnetwork

Table 1 summarizes the different FER methods in terms of accuracy, model complexity and underlying strategy. The comparison highlights that while data augmentation and full model fine tuning can improve performance, they still rely on adapting the entire network, which may lead to inefficient parameter utilization.

3 Methodology

The proposed methodology formulates model design as a structured subnetwork selection problem within a fixed pretrained backbone, inspired by Neural Architecture Search (NAS) principle [23] and path-based network selection, with the primary objective of identifying an optimal subnetwork within the fixed MobileNetV2 backbone for facial expression recognition [16].

Unlike PathNet, which relies on evolutionary strategies for pathway optimization, our approach adopts a deterministic and structured search strategy tailored to a single task FER setting. Specifically, we restrict the search space to the subnetwork where candidate subnetworks are formed by selecting the consecutive blocks from the MobileNetV2 architecture.

To efficiently evaluate the candidate subnetwork, we employ a linear probing strategy [24], where features are extracted from the frozen pretrained layers and evaluated using a lightweight classifier. This enables rapid comparison of subnetwork without requiring full end to end training. MobileNetV2 is chosen as the backbone due to its lightweight design, making it well-suited for environment where the resource is constrained and hence align well with the goal of parameter efficient learning under limited data conditions.

3.1 Problem Formulation

Let MobileNetV2 be composed of a sequence of sequential blocks, instead of utilizing the full network, we consider a subnetwork formed by selecting the first k blocks, where k controls the number of blocks.

Given a limited labeled dataset, the objective is to identify the most optimal subnetwork that provides the best tradeoff between model capacity and generalizations that are evaluated using a lightweight classifier.

The problem can hence be formulated by selecting the subnetwork that maximizes the classification performance under limited data constraints. This enables efficient exploration of model capacity without requiring full network training.

3.2 Model Architecture

The backbone of this method is the MobileNetV2, [25] which is a lightweight model designed to work well on mobile phones and resource constrained devices, while also maintaining strong performance on image classification tasks. The architecture consists of an initial convolutional layer, which is followed by a sequence of inverted residual blocks and a final convolutional layer that produces high level feature representations.

MobileNetV2 consist of an initial convolutional block , followed by 17 inverted residual blocks and a final convolutional layer. Each inverted residual block contains a depth wise separable convolution with an expansion layer an a linear bottleneck. In this work, block indices refer exclusively to these 17 inverted residual blocks. Each candidate subnetwork consists of the initial convolutional block, the first k inverted residual blocks, the final convolutional layer and a global average pooling layer, followed by a linear classifier. This structure ensures that all candidate subnetworks produce a fixed length feature vector regardless of the truncation depth k. The model is inherently sequential, where earlier blocks capture low-level features such as edge and textures, and later blocks capture increasingly abstract representations. This hierarchy justifies restricting the search space to prefix subnetworks, as removing later blocks removes the most abstract and potential dataset specific features first.

3.3 Search Strategy

In this work, we formulate the architecture design problem as a subnetwork selection task within the pretrained model. Rather than training the full network model, we aim to identify a subset of layers that achieves a strong performance under the limited data condition. The search space consists of a sequence of inverted residual blocks, where each subnetwork will include blocks from layer 1 up to layer k, with $k \in \{3 \dots \dots 17\}$. This design will allow the exploration of different network depths while maintaining structural consistency. The search space is restricted to prefix subnetworks rather than arbitrary subsets because MobileNetV2 is designed as a sequential feature hierarchy, where earlier blocks capture low level features and later blocks capture increasingly abstract representations. Prefix selection preserves this hierarchy while allowing structured depth exploration.

To efficiently evaluate each candidate, we have adopted a linear probing approach. The convolution backbone will be kept frozen, and feature representations are extracted once for both training and testing subsets. A logistic regression classifier is then trained on these features to estimate the classification accuracy. This avoids end-to-end training for each candidate architecture, hence significantly reducing computational cost.

To ensure that a fair comparison, the search phase is conducted as a completely independent experiment. A fixed 15% subset of the training set is selected for efficiency and further split into training 80% and validation 20% subsets and the test set is not used during the search phase. Features are cached to further reduce the computational overhead. To assess the robustness of the search, the process is repeated across multiple random seeds. The results identified the blocks 1-13 as the optimal subnetwork in four of five seeds, demonstrating the stability and reliability of the proposed search strategy.

Compared to Neural Architecture Search (NAS) approaches, the proposed method provides significantly more efficient and deterministic alternative, making it suitable for low data and resource constrained settings.

3.4 Training Strategy

To evaluate the performance of the proposed subnetwork under both standard and low data conditions, experiments are conducted on two datasets: FER and CK+.

For the FER2013 dataset, a limited data regime is simulated by training the models on different fractions of the training set. Specifically, we consider fractions of 5%, 10%, 20%, 50% and 100% of the available data. For each fraction, a subset of the training data is sampled in a deterministic manner based on the random seed. To ensure robustness, each experiment is repeated across five different seeds, resulting in a total of 25 training configurations. Unlike search phase, the training experiments use only training sets. An internal 80/20 split of the training data is used for early stopping and the full held-out test is used exclusively for final evaluation. The test is never seen during training or subnetwork selection.

We compared two models, first is a full MobileNetV2 baseline and second is the proposed parameter efficient subnetwork achieved during the search phase. Both models are initialized with ImageNet pretrained weights. All experimental settings, including optimization, data splits and augmentation are kept identical between baseline and evolved model to ensure a fair and controlled comparison.

During the training, early stopping is applied with a patience of 7 epochs to prevent overfitting, which is particularly critical in low-data settings. Additionally, for both the models, the initial convolutional block [0] is frozen during training, while all remaining blocks are trainable, ensuring a symmetric comparison. Both the models are trained with a label smoothing of 0.1, a weighted random sampler to address class imbalance and an internal validation split for early stopping. A dropout of 0.2 is applied for the classifier for FER2013 and 0.4 for CK+, reflecting the difference in dataset size. Standard data augmentation techniques are applied such as random horizontal flipping and rotation during training to improve generalization. All models are evaluated on the full tests set without augmentation.

For the CK+ dataset, a subject independent evaluation protocol [4] is used to prevent identity leakage. A 5-fold, cross validation method is used at subject level, which ensures that all samples from a given subject appear exclusively in either the training or testing set. Within each fold, the non-test subjects are further divided at the subject level into a training set (85%) and an internal validation set (15%). The validation set is used for early stopping and checkpoint selection, while the test fold is kept completely unseen until final evaluation. The cross-validation process is repeated across three random seeds, resulting in a 15-evaluation run per model. The architecture selected during search phase is fixed and used across all folds and seeds. Specifically, the subnetwork identified on FER2013 consisting of the initial convolutional block, inverted residual blocks 1 through 13, the final convolutional layer and global average pooling is used directly for CK+ training without any further architecture modification.

To assess the statistical significance, paired t-tests and Wilcoxon signed rank tests are performed between the baseline and proposed model across the seeds for each data fraction. This will help in comparison with performance differences under different data conditions.

4 Results

4.1 Architecture Search Result

An exhaustive search was conducted on the FER 2013 dataset to identify the most optimal blocks from MobileNetV2. Figure 1 illustrates the relationship between classification accuracy and network depth. The accuracy steadily improved from 37.98% at block 1-3 and peaks at 1-13, after which performance begins to decline.

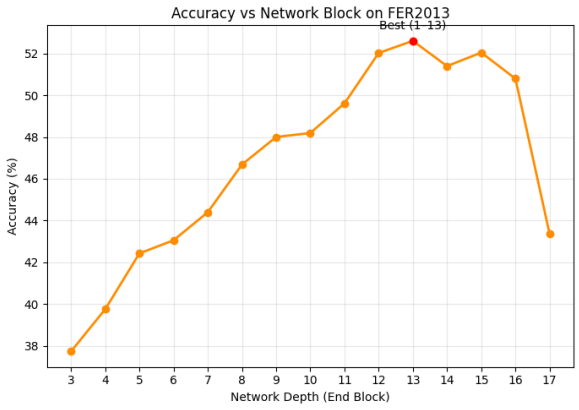

Fig. 33. Accuracy vs Network Block

The best performing block consists of blocks 1-13, achieving a mean accuracy of 53.11% ± 0.58% while reducing the parameter count to 0.99M, corresponding to 55.8% reduction compared to full model. This trend indicates that the intermediate layers capture the most discriminative features for facial expression, while deeper layers may introduce redundancy or overfitting.

4.2 Stability Analysis

Table 13. Top 3 performing network blocks on FER2013 dataset.

Blocks	Mean(%)	Std(%)	wins	S42(%)	S123(%)	S456(%)	s789 (%)	S999(%)
1-13	53.11%	0.58	4/5	52.65	53.44	53.30	52.28	53.90
1-12	51.94%	0.30	1/5	51.76	52.00	51.49	52.37	52.09
1-15	51.69%	0.44	0/5	51.58	52.07	51.21	52.32	51.28

To evaluate the robustness of the search results, the search was repeated across five seeds using a fixed 15% subset of the training set, yielding a mean accuracy of 53.11% for blocks 1-13. The purpose of this experiment was to verify whether the optimal architecture identified in the single seed search remains consistent under different classifier initialization.

Table 2 shows the top three performing subnetworks. From the table it shows that the subnetwork consisting of blocks 1-13 achieves the highest accuracy in four of five seeds, with accuracies of 52.65%, 53.44%, 53.30%, 52.28% and 53.90% respectively. The corresponding mean accuracy of 53.11% with a standard deviation of 0.58% indicates low variance and high stability. Blocks 1-13 achieves 4/5 wins, with seed 789 selecting blocks 1-12 as a marginal alternative (52.37% vs 52.28%), a difference of only 0.09%, confirming that blocks 1-13 remains the most reliable blocks.

In contrast the deeper configuration had slightly lower mean performance despite having high parameter count, further supporting the effectiveness of the proposed subnetwork selection strategy.

4.3 Limited Data Results

Table 3 presents the performance of the proposed subnetwork under limited data conditions across all data fractions; the subnetwork achieves a test accuracy within 0.5% of the full mobileNetV2baseline at 10% and 20% data fractions. At these scales the difference is not significant, indicating the two models are statistically equivalent despite a 55.8% parameter reduction. At 5% ,50% and 100%, the full model achieves significant advantage of 2.89%, 0.81% and 1.72% respectively suggesting that the full model benefits from higher representational capacity at data extremes. Across all the fractions the subnetwork exhibited a train test accuracy gap of between 35% to 48% lower than the baseline depending on data fraction, indicating substantially reduced overfitting regardless of data scale.

Table 14. Performance comparison under limited data conditions on FER2013

Fractions	Baseline	Evolved	Diff	p	Baseline Gap	Evolved Gap
5%	49.27%±0.53%	46.38%±1.66%	-2.89%±1.54%	0.0198	44%	23%
10%	52.84%±0.56%	52.32%±0.29%	-0.52%±0.65%	0.1857	41%	23%
20%	57.55%±0.33%	57.14%±1.18%	-0.41%±1.37%	0.5810	38%	22%
50%	62.73%±0.41%	61.92%±0.15%	-0.81%±0.54%	0.0393	32%	21%
100%	67.00%±0.24%	65.27%±0.27%	-1.72%±0.43%	0.0013	29%	19%

4.4 Subject Independent Evaluation of CK+

Under the subject-independent 5-fold cross validation protocol repeated across three random seeds, the evolved model achieved 90.47% ± 3.65% , compared to 88.58% ± 3.06% a difference of +1.89% ($t = 1.607$, $p = 0.1303$; Wilcoxon $W = 23.0$, $p = 0.1159$). While this difference did not reach a conventional significance threshold, per class analysis shows the evolved model improving on 6 of 7 emotion categories by F1 score. The most notable gains were in contempt (F1:0.659 to 0.726) and sadness (F1:0.727 to 0.780). Macro F1 improves from 0.834 to 0.862.

Table 15. Per-class F1 scores on CK+ (baseline vs Evolved, aggregated across 3 seeds x 5 folds)

Emotion	Baseline F1	Evolved F1	Difference
Anger	0.784	0.812	+0.028
Contempt	0.659	0.726	+0.067
Disgust	0.915	0.942	+0.027
Fear	0.837	0.840	+0.003
Happy	0.946	0.951	+0.005
Sadness	0.727	0.780	+0.053
Surprise	0.973	0.981	+0.008
Macro F1	0.834	0.862	+0.028

4.5 Discussion

This study was able to demonstrate that the proposed subnetwork strategy achieves parameter efficient FER through deterministic, computationally lightweight search procedure. The selected subnetwork reduces the model complexity by 55.8% while maintaining comparable accuracy at moderate data regimes and on CK+, without statistically significant differences compared to the full mobileNetV2 baseline. Additionally, it demonstrated reduced overfitting, as indicated by a smaller train test gap across most evaluated settings.

A key insight is that removing the later blocks results in competitive accuracy on CK+, while consistently reducing overfitting across all conditions, suggesting that the deeper layer of the full model introduces features that do not generalize well under limited data.

By selecting a more compact subnetwork, the model retains the most information regarding the subject while discarding the unnecessary complexity. This aligns with the broader principle that

increased blocks does not always translate to better generalization, especially in constrained or domain specific datasets.

The stability analysis confirms the reliable search behavior, with block 1-13 selected in four of five seeds, with single disagreement being a margin of 0.09% difference in favor of blocks 1-12. The limited data experiments show that the compact subnetwork achieves equivalent accuracy to the full model in the 10-20% data range, while its train test gap remains between 35% and 48% lower than the baseline depending on data fraction, suggesting the compact architecture provides structural regularization independence of data scale.

An important contribution to this work is the efficiency gain. The evolved model achieves 55.8% reduction in parameter while maintaining competitive accuracy and improving generalization behavior. This has practical implications for deployment in a resource constrained environment, where both computational efficiency and predictive accuracy are critical.

Despite these promising results, several limitations should be acknowledged. The experiments are conducted on a relatively limited dataset, which may restrict the generalizability of the findings to larger or more diverse scenarios. Additionally, the search space was constrained to predefined block truncation, and more flexible or adaptive architecture might be able to yield further improvements.

4.6 Conclusion

In this work, a subnetwork selection approach was proposed to improve the performance and efficiency of a baseline MobileNetV2 model for facial expression recognition under limited data conditions. Through comprehensive evaluation including architecture search, stability analysis, limited data experiments and subject independent cross validations, the evolved architecture achieved performance comparable to the baseline at 10-20% data fractions, with reduced overfitting across all evaluated conditions and competitive accuracy of 90.47% on CK+ which is +1.89% higher than the baseline.

The results show that the reduced model configuration achieves a 55.8% reduction in parameters while maintaining competitive accuracy within 1-3% of the full MobileNetV2 baseline. These findings demonstrate that substantial model compression can be achieved within this experimental settings, with statistically significant differences observed only at data extremes, where the full model retains an advantage due to higher representational capacity.

These findings highlight that significant parameter reduction can be achieved without statistically significant loss in performance while improving class balanced generalization.

4. References

[1] T. a. S. V. a. V. N. a. D. P. Kopalidis, "Advances in Facial Expression Recognition: A Survey of Methods, Benchmarks, Models, and Datasets," *Information,* vol. 15, no. 3, 2024.

[2] D. L. Bin Li, "Facial expression recognition via ResNet-50," *International Journal of Cognitive Computing in Engineering,* vol. 2, pp. 57-64, 2021.

[3] W. S. E. I.Michael Revina, "A Survey on Human Face Expression Recognition Techniques," *Journal of King Saud University - Computer and Information Sciences,* vol. 33, no. 6, pp. 619-628, 2021.

[4] S. a. D. W. Li, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing,* 2022.

[5] Y. K. a. Z. Chen, "Facial Emotion Recognition: State of the Art Performance on FER2013," 2021.

[6] P. C. B. a. P. P. Barile, "Transfer Learning for Facial Emotion Recognition on Small Datasets," *Journal of Systemics, Cybernetics and Informatics,* vol. 22, no. 4, pp. 1-5, 2024.

[7] F. a. Q. Z. a. D. K. a. X. D. a. Z. Y. a. Z. H. a. X. H. a. H. Q. Zhuang, "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE,* vol. 109, no. 1, pp. 43-76, 2021.

[8] C. A. S. A. S. a. E. H. R. A. A. Istiqomah, "Facial Expression Recognition using Convolutional Neural Networks with Transfer Learning ResNet-50," *Journal of Applied Informatics and Computing,* vol. 8, no. 2, pp. 257-264, 2024.

[9] S. M. B. N. Kiran Maharana, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings,* vol. 3, no. 1, pp. 91-99, 2022.

[10] C.-L. K. B.-G. Kim, "Few-shot learning for facial expression recognition: a comprehensive survey," *Journal of Real-Time Image Processing,* vol. 20, 2023.

[11] D. N. D. T. S. S. D. S. N. T. T. D. D. F. C. Nguyen, "Meta-transfer learning for emotion recognition," *Neural Computing and Applications,* vol. 35, no. 14, pp. 10535-10549, 2023.

[12] G. K. a. P. J. a. D. S. K. a. S. P. Sahoo, "Deep Leaning-Based Facial Expression Recognition in FER2013 Database: An in-Vehicle Application," in *2022 IEEE 19th India Council International Conference (INDICON)*, 2022.

[13] I. J. G. a. D. E. a. P. L. C. a. A. C. a. M. M. a. B. H. a. W. C. a. Y. T. a. D. T. a. D.-H. L. a. Y. Z. a. C. R. a. F. F. a. R. L. a. Xi, "Challenges in Representation Learning: A report on three machine learning contests," 2013.

[14] P. a. C. J. F. a. K. T. a. S. J. a. A. Z. a. M. I. Lucey, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010.

[15] Y. a. Z. C. Yu, "Facial Expression Recognition Based on Improved MobileNetV2 Network Model," in *2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT)*, 2024.

[16] M. Y. a. C. G. a. L. N. a. D. Z. a. A. K. a. Q. Liu, "Good Subnetworks Provably Exist: Pruning via Greedy Forward Selection," 2020.

[17] H.-W. a. N. V. D. a. V. V. a. W. S. Ng, "Deep Learning for Emotion Recognition on Small Datasets using Transfer Learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*, Seattle, Washington, USA, 2015.

[18] S. Q. X. L. Z. M. a. J. L. Y. Zhong, "Facial Expression Recognition Based on Optimized ResNet," in *2020 2nd World Symposium on Artificial Intelligence (WSAI)*, Guangzhou, China, 2020.

[19] S. Deng, Y. Sun and E. Galvan, "Neural Architecture Search Using Genetic Algorithm for Facial Expression Recognition," 2022.

[20] W. Z. a. A. G.-C. a. V. S. a. Y. L. a. N. Carion, "Light-weight probing of unsupervised representations for Reinforcement Learning," in *Proceedings of the Reinforcement Learning Conference (RLC 2024)*, online, 2024.

[21] M. S. a. A. H. a. M. Z. a. A. Z. a. L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2019.