Master-Thesis in Biomedical Engineering
Fakultät Mechanical and Medical Engineering

Decoding Emotions

How temporal modelling enhances recognition accuracy

Sreerag Chandrasekharan

Matriculation Number: 272750

Email: sch41871@hs-furtwangen.de

Supervisors:

Prof. Dr. Knut Möller

Herag Arabian

Villingen-Schwenningen, 28th. March 2024

Abstrakt

Facial emotion recognition (FER) is pivotal in human-computer interaction and can be used for therapies for individuals with autism spectrum disorder (ASD). Deep learning techniques offer promising avenues for improving FER, but achieving high accuracy and rapid response times remains a challenge. This study introduces two novel models integrating temporal modeling to enhance accuracy by exploiting time related dependencies.

These models employ pre-trained networks like ResNet50, GoogleNet (Inception), and AlexNet. The recognition accuracy of the models was evaluated subsequent to integrating LSTM with the pre-trained networks. The study uses the Oulu-CASIA database and employs techniques like cascade face detection, feature extraction, and data augmentation to enhance model performance. Also, compared the performance between two models with different data inputs.

The models' effectiveness is evaluated by repeated training and validation. Model 1, trained on augmented datasets, shows improved accuracy, especially with noise and brightness variations in the training data. Moreover, Model 2, with LSTM architecture, consistently outperforms Model 1 across all pre-trained models. GoogleNet with LSTM achieves significant accuracy improvement from 83.51% to 94.21%, while AlexNet and ResNet50 also exhibit notable enhancements with LSTM.

In conclusion, this research underscore the importance of LSTM architecture in FER tasks, leading to substantial accuracy improvements across deep learning models. The integration of LSTM networks offers a promising way for future research, suggesting further improvements in real-time emotion recognition systems.

Erklärung zur Abschlussarbeit

Hiermit erkläre ich ausdrücklich, dass ich, Sreerag Chandrasekharan, die eingereichte Diplomarbeit selbstständig und ohne fremde Hilfe angefertigt habe. Ich habe außer der zitierten Literatur und anderen in der Arbeit erwähnten Quellen keine externe Unterstützung in Anspruch genommen. Sämtliche Literatur und sonstige Quellen, die ich bei der Erstellung dieser wissenschaftlichen Arbeit wörtlich oder inhaltlich herangezogen habe, habe ich deutlich gekennzeichnet und gesondert aufgeführt.

Declaration of Master Thesis Statement

I hereby formally declare that I, Sreerag Chandrasekharan have written the submitted thesis independently and without unauthorized outside help. I did not use any outside support except for the quoted literature and other sources mentioned in the paper. I clearly marked and separately listed all of the literature and all of the other sources which I employed when producing this academic work, either literally or in content.

Villingen-Schwenningen, 28th. March 2024

Addresse : Salinenstße 46, Schwenningen, 78054

Unterschrift/Signature

Acknowledgement

I extend my sincere gratitude to my Supervisor, Prof. Dr. Knut Möller, for giving me the opportunity and for his trust in me.

Sincerely thanks to my second supervisor, Herag Arabian, for believing in me and allowing me to work on my master thesis. I am grateful for his valuable assistance in providing me with the resources I required to make the right decision and successfully complete my thesis. I am deeply indebted to him for the opportunity to have him advise me in projects during my academic pursuit in MSc. Biomedical Engineering.

I am extremely grateful for providing me with the resources and for the perspective feedback that I have received, which has helped me sharpen my skills and to bring to fruition this project. I am deeply filled with gratitude to my supervisors for their constant guidance.

Contents

List of Figures

Chapter 1

Introduction

1.1 Facial emotion recognition

Facial emotion recognition (FER) is the process of detecting basic emotions such as fear, happiness, sad, disgust, anger and surprise through the analysis of human's facial expressions [16] rr[**chandrasekharan2024decoding**]. This is an important part of human-computer interaction and can be applied in the healthcare sector for assessing therapeutic treatments for autism spectrum disorders [3] [9].

Achieving precision and stability in FER through computer algorithms is a complex task, impacted by the difference in human facial structures and image quality, such as changes in facial orientation and illumination [16]. In the field of FER, it is observed that convolutional neural networks (CNNs) and other deep learning frameworks are particularly effective [16].

CNNs are networks employed in tasks related to classification and computer vision, such as categorizing images and identifying objects [13].CNNs are structured with various layers, comprising an initial input layer, multiple intermediate hidden layers, and a final output layer. CNNs stand out from other neural networks due to their exceptional performance when dealing with inputs such as images, speech, or audio signals [13].

1.2 Deep learning influence on FER

Deep learning has a substantial impact on FER. The progress in deep learning has resulted in enhanced accuracy in facial detection in real time [7]. Numerous effective methods have been established for FER utilizing deep learning techniques [18, 29, 37, 42, 14].

Ko in 2018, carried out a concise study of FER based on visual data. This research traced the progression from traditional FER methods to those based on deep learning. The study highlighted the emergence of new hybrid models that combine CNNs for analyzing spatial characteristics and long short-term memory (LSTMs) for examining temporal features [18].

Additionally, a study was carried out by Park et al. by utilizing a 3D convolutional neural network based on a multirate scheme. The research employed intensity-based normalization for the input images prior to classification. research demonstrated an accuracy of 99.79% when tested on the GEMEP-FERA database [29].

In 2021, Verma et al. introduce AutoMER, an innovative algorithm for conducting spatiotemporal architecture for the recognition of microexpressions (MER). This method employs an original parallelogram-shaped search space and implements a unique 3-D singleton convolution technique for analyzing data at the cellular level.The effectiveness of AutoMER is evidenced by its superior performance over current leading methods across five different microexpression datasets such as CASME-I, SMIC, CASME-II, and SAMM [37].

Also, Zheng et al. have developed a novel hybrid neural network model designed to enhance real-time recognition of micro-expressions by using CNN and recurrent neural network (RNN), like LSTMs. The system is engineered to process brief facial video clips and accurately identify the micro-expressions with the help of facial micro-expression datasets that are readily accessible to the public for both training and evaluation purposes. This research further introduce metrics for score fusion and improvement within their study. Their proposed hybrid model outperforms literature-reported methods tested on the same datasets [42].

A study carried out by Jiang et al. introduce a groundbreaking framework utilizing deep neural networks for objectively determining the presence of major depressive disorder (MDD) through video analysis. This system was proposed for recognizing facial expressions, using CNN and trained on a comprehensive, openly accessible dataset. This method was tested using 365 video interviews of 12 individuals suffering from depression. This study show the possibility of using passively collected video data to classify remission from MDD and the response to Deep Brain Stimulation treatment [14].

In this study, three different pre-trained CNNs - ResNet [11], GoogleNet [35], and AlexNet [20] - are utilized. These CNNs are then incorporated with a LSTM network to examine the impact on the recognition accuracy of the models. The Oulu-CASIA [40] dataset serves as the basis for this investigation. The integration of LSTM is used to determine the capability of the models to capture temporal dependencies.

1.3 Temporal Modeling influence on FER

Temporal modeling in FER determine how facial expressions change over time. It's a process in which the alterations in facial features and expressions are extracted by using a series of video frames. LSTM networks, a type of recurrent neural network (RNN), are used for sequence learning methods [31]. Emotion sequences with different numbers of frames are utilized as an aspect of temporal modeling in this study. Several research have focused on determining FER through the integration of CNNs and temporal modeling [33, 42, 8, 39, 34].

Recently, Singh et al. proposed a model that merges 3D-CNN and Convolutional LSTM (ConvLSTM). The conventional fully-connected LSTM (FC-LSTM) converts the image into a one-dimensional vector, leading to a loss of vital spatial information. To overcome this, ConvLSTM was introduced, which can perform LSTM operations in convolutions without altering the original state of the image. This approach is capable of capturing spatiotemporal information from video sequences of emotions. This model achieved competitive recognition accuracy using three publicly available FER datasets such as SAVEE, CK+, and AFEW [33].

Additionaly, Zheng et al. developed a model that integrates a CNN, LSTM, and a

Vision Transformer for the purpose of real-time micro-expression recognition. The CNN component is responsible for spatial feature extraction within an image, while the LSTM component focuses on summarizing temporal features. The Vision Transformer, equipped with an attention mechanism, is able to identify sparse spatial relations either within an image or in video frames. This neural network models were put to the test using facial micro-expression datasets, which allowed for the recognition of various micro-expressions such as happiness, fear, anger, surprise, disgust, and sadness. The application of score fusion significantly enhanced the recognition accuracy of this model [42].

With an emphasis on landmark-based and image-based techniques, Farkhod et al. carried out a comparative study of the efficacy of live testing methods for face expression recognition among people wearing masks in 2022. The key steps of the proposed method involve face detection using the Haar-Cascade technique, landmark identification utilizing a MediaPipe face mesh model, and training the model on seven emotional categories. The FER-2013 dataset was utilized for model training. The findings indicate that the proposed model attains an overall accuracy of 91.2% across seven emotional classes in image-based applications. However, the accuracy is relatively lower for real-time emotion detection [8].

Zhang et al. in 2017 proposed a facial expression recognition method based on deep evolutionary spatial-temporal networks. The methodology comprises two networks: a part-based hierarchical bidirectional recurrent neural network (PHRNN) designed to capture the dynamic changes in facial physical structure from video data, and a multi-signal convolution neural network (MSCNN) aimed at extracting spatial features from individual frames. The PHRNN is tasked with modeling facial morphological variations and the dynamic evolution of expressions, while the MSCNN supplements this information with still appearance features. Training of these networks utilized datasets including CK+, Oulu-CASIA, and MMI. The PHRNN attained an accuracy of 96.36% on the CK+ dataset and 78.96% on the Oulu-CASIA dataset. [39].

Furthermore, Sun et al. introduced a dynamic sequence facial expression recognition system that integrates both shallow and deep features utilizing attention mechanisms. Shallow features are obtained from facial landmarks and local texture characteristics following the facial action coding system (FACS), whereas deep features are extracted employing an enhanced AlexNet structure. BiLSTM is incorporated to consolidate feature relationships across frames, thereby augmenting facial expression recognition for real time. The system attains accuracies of 99.1%, 89.88%, and 87.33% on CK+, MMI, and Oulu-CASIA databases, respectively [34].

1.3.1 Recurrent neural network

Recurrent neural networks (RNNs) possess cyclic connections, rendering them more adept at modeling sequence data compared to feedforward neural networks. They have exhibited significant efficacy in tasks involving sequence labeling and prediction, such as handwriting recognition and language modeling [32].

In particular, Long Short-Term Memory (LSTM), a variant of RNNs, has proven to be highly successful in numerous sequence prediction and labeling endeavors. Two main variations of LSTM exist, namely:

1. Conventional LSTM: This design comprises memory blocks housing memory cells and gates for regulating information flow. The gates incorporate an input gate, an output gate, and a forget gate. The input gate oversees the control of input activation entering the memory cell, the output gate governs the release of cell activation, and the forget gate modulates the internal state of the cell. This architecture enables LSTM to effectively retain long-term temporal contextual information [32].

2. LSTMP (LSTM with recurrent projection layer): The LSTMP architecture represents a refinement of the traditional LSTM structure. It introduces a distinct linear recurrent projection layer positioned after the LSTM layer. The recurrent connections now extend from this recurrent projection layer to the input of the LSTM layer, while the network output units are linked to this recurrent layer. This design modification serves to diminish the parameter count and alleviate the computational demands associated with training LSTM models [32]. The difference in LSTM architecture are shown in Figure 1.1.

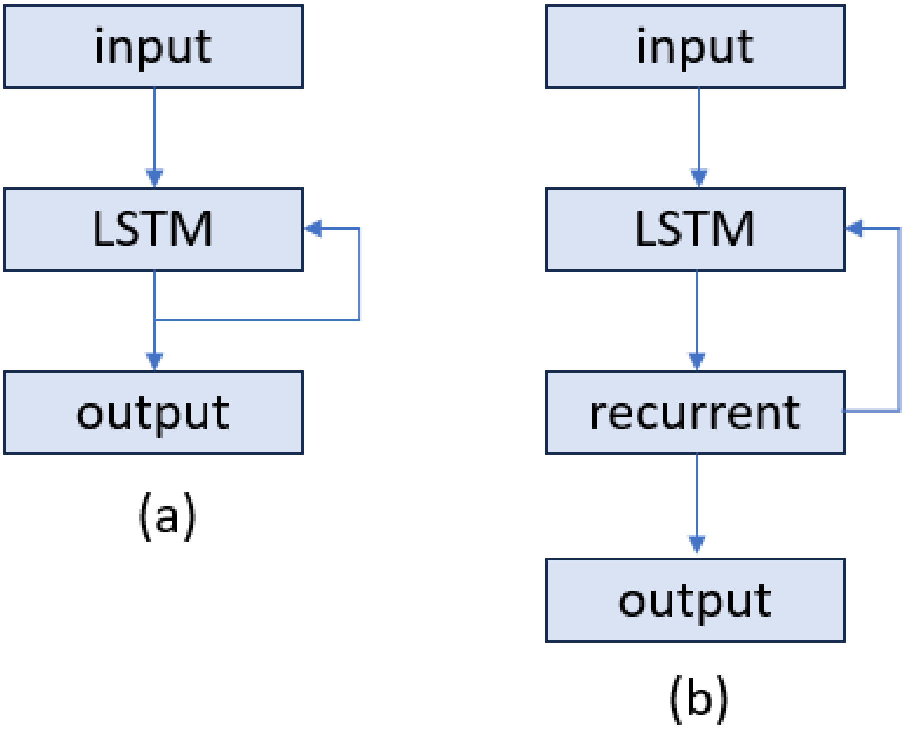


Figure 1.1: (a) Conventional LSTM architecture, (b) LSTMP architecture

1.4 FER impact on autism spectrum disorders (ASD)

Autism spectrum disorders (ASD) is type of disorder identified in children at approximately 3 years of age [24]. The identification of emotional states in individuals with autism is important for parents and caregivers, as it allows them to provide care and assistance for their needs [36]. In response to this concern, two models were developed with the aim of discerning the emotional states of individuals through FER.

Takahashi et al. explored facial expression recognition, focusing on individuals with ASD using predictive processing framework. They employed recurrent neural networks (RNNs) to model the dynamic changes in facial expressions related to six primary emotions. The study aimed to mimic developmental learning by not relying on explicit emotion labels. The effectiveness of these networks was then evaluated based on their capacity to recognize unfamiliar facial expressions [36].

This investigation prioritized on the development and comparison of two models using the OULU-CASIA [40] dataset. The first model, known as Model 1, was constructed using a transfer learning approach with pre-trained network architectures. The second model, known as Model 2, was built using the features extracted from Model 1 after its training. Model 2 integrated LSTM layers and utilized the features derived from the final global average pooling layer of the network architecture used in Model 1 for its training.

The study incorporates three different input types for training: original images, images containing noise, and images with both noise and increased brightness. The Oulu-CASIA [40] dataset is utilized to evaluate the models' performance. Pretrained networks are augmented with LSTM layers to explore potential improvements in recognition accuracy. Following this, the datasets of the trained models are utilized to compare classification accuracy before and after LSTM integration, and to evaluate model performance across various input data variations.

This study aims to investigate the effectiveness of RNN like LSTM, in improving the accuracy and robustness of facial emotion recognition systems. The objective of this study is to explore the impact of FER on ASD, highlighting the potential benefits of utilizing FER technology in assisting individuals with ASD.

Chapter 2

Material and methods

2.1 Pretrained networks

Pre-trained networks are models that have been previously trained for another task or large datasets.The features and patterns learned by such networks can be used for similar tasks [5]. They can greatly shorten the training period and improve a model's efficiency, especially when there is limited input data [5]. Pretrained network that used for this study are mentioned below

- ResNet50: The ResNet50 [11] employs residual mapping to address the issue of saturation degradation. Pre-trained ResNet50 models are capable of capturing complex facial features [12] .

- GoogleNet (Inception): GoogleNet [35], represents a convolution neural network design recognized as Inception. It uses inception modules to efficient computational resources while maintaining a large receptive field. This can be beneficial for capturing different scales of facial expressions [2].

- AlexNet: AlexNet [20] is one of the pioneer deep learning network. It has five convolution layers and three fully connected layers. It incorporates convolutional layers, max pooling layers, and dense layers as its fundamental components [19].

2.2 Optimization of Model

The optimization process involves fine-tuning the model architecture by reducing unnecessary connections or parameters [30]. Stochastic gradient descent with momentum (SGDM) and Adaptive moment estimation (Adam) are the two optimization technique used in this study for tuning the models architectures.

1. SGDM: SGDM is a variant of Stochastic gradient descent. It introduces momentum to accelerate convergence by adding a fraction of the previous update to the current one. SGDM processes one sample at a time, updates the weights accordingly [23]. The SGDM solver was used to train the LSTM network model.

2. Adam: In Adam optimization technique, it employs an exponential weighted average approach and combination of momentum [17]. This method is utilized during the training of Model 1.

2.3 Scheduling of learning rate

The learning rate is the rate at which weight modifications occur throughout the training phase. The general strategy is to employ a relatively high learning rate during the initial stages of training to facilitate rapid convergence, and then reduce the learning rate towards the end of the training to allow the network to converge more accurately to the optimal solution [22].

In this research, the learning rate is configured to 0.0001 to prevent bypassing local minima.

2.4 Transfer learning

This method is highly practical and efficient for tasks with limited data availability. The process of transfer learning involved training the convolution neural network model using a substantial amount of data [2]. In the preparation of this model, it undergoes two phases, fine-tuning phase to train on Oulu-CASIA dataset and feature extraction phase.

2.4.1 Fine-tuning

Fine-tuning is a process that takes a pre-trained model and "tunes" it for a different but related problem. This uses weights from a pre-trained model as the initialization for a new model [25]. The following outlines the procedures employed for fine-tuning in this study.

1. Pretraining: Pretrained network such as GoogleNet [35], AlexNet [20] and ResNet50 [11] was used for training with the help of Oulu-CASIA dataset.
2. Creating the model: A new neural network model is created by using the architecture and parameters of the pretrained Networks, except for the output layer.
3. Adding the Output Layer: The target model is enhanced with an output layer. The number of outputs is determined as seven, which corresponds to the total number of emotion classes in the target dataset of Oulu-CASIA.
4. Training the Model: The target model is then trained on Oulu-CASIA dataset. The output layer's parameters are newly learned, while the other layers' parameters are fine-tuned based on the parameters of the pretrained networks.

2.4.2 Feature extraction

Feature extraction is the process of converting raw input, often in the form of static or dynamic image, into numerical attributes which is processed by machine learning models [41]. The technique employed for feature extraction from emotional sequences from pre-trained models such as GoogleNet [35], AlexNet [20], and ResNet50 [11] as shown in figure 2.1. Specifically, the features are derived from the *pool5-7x7_s1* layer of GoogleNet, the *avg_pool* layer of ResNet50, and the *relu7* layer of AlexNet.

2.5 Data pre-processing

Utilizing the cascade detection algorithm developed by Viola and Jones in 2001, the database was modified to remove any extraneous background artifacts [38]. A method incorporating local binary patterns was employed for the selection of ap-

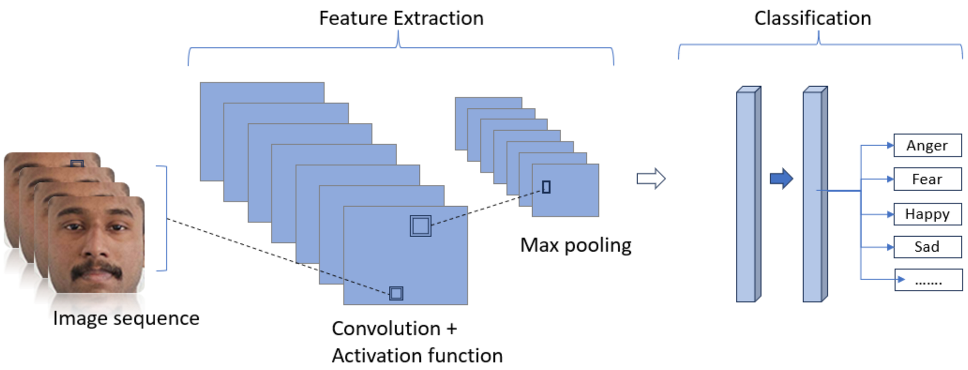


Figure 2.1: Process of feature extraction from image sequence

propriate faces for the training phase, mirroring the technique discussed by Arabian et al. in 2021 [3]. Before proceeding with the training and validation phases, these images were deliberately subjected to various types of noises.

2.5.1 Cascade face detection

Cascade detection represents a method within machine learning process for the identification of visual objects, particularly effective in recognizing faces [38]. This approach stands out for its ability to process images faster and a high level of accuracy in detection [38]. It employs a series of classifiers to quickly assess regions within an image to differentiate face and non face region [27]. The cascade face detection method is described as follows:

1. Feature identification: The technique employs Haar features, essentially rectangular patterns, to identify facial characteristics as shown in figure 2.2. The characteristics including borders, textures, and contours are used for detection. The patterns are varied in size and placed in different positions over the image to generate a broad collection of potential features for discrimination [27].

2. Adaboost Filtering: Adaboost is one of the important features that help distinguish between facial and non-facial areas. Through iterative training of simple classifiers, each focused on a single feature out of large number of features. For each features, the optimal threshold is identified, which classifies

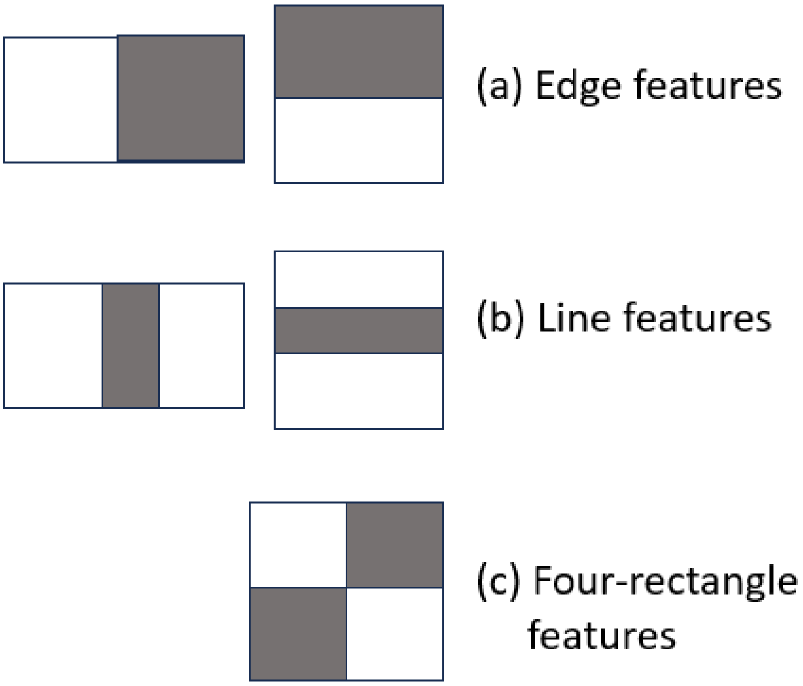


Figure 2.2: Haar features used for Cascade face detection.

into face and non face region [27].

3. Classifier Cascade: This approach does not assess all attributes on each image segment but instead utilizes a cascade of classifiers. Each sub-window of the cascade uses a selection of features; if an image segment doesn't pass a sub-window, it's instantly eliminated, thus conserving computational resources. Only those segments that succeed at all sub-windows are deemed possible facial areas. The method of classifier cascade are represented in figure 2.3 [27]

2.5.2 FrontalFaceLBP

The FrontalFaceLBP technique is a method for face recognition within the field of computer vision. Predominantly, face detection employs the Viola-Jones algorithm [38], which is predicated on the utilization of Haar features [27]. This method is efficient in identifying faces that are positioned upright and looking straight faces.

In contrast, employing the local binary pattern (LBP) method proved to be more

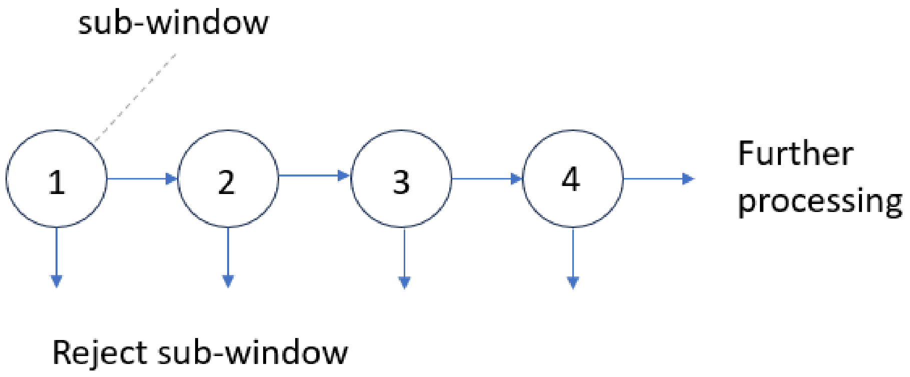


Figure 2.3: Image processing through classifier cascade, sub-windows are numbered as different levels.

efficient in detecting faces, particularly those oriented upright and directly facing the camera [28]. The robustness of LBP features in adapting to variations in lighting conditions underscores its effectiveness [26].

Local binary patterns

This method is used for extraction of face from an raw images in the Oulu-CASIA [40] dataset. Figure 2.4 illustrates the extraction of face from image using image processing technique FrontalFaceLBP. This method effectively isolates the facial region by eliminating all other distractions. These processed images are then resized to be fed into the network architecture.

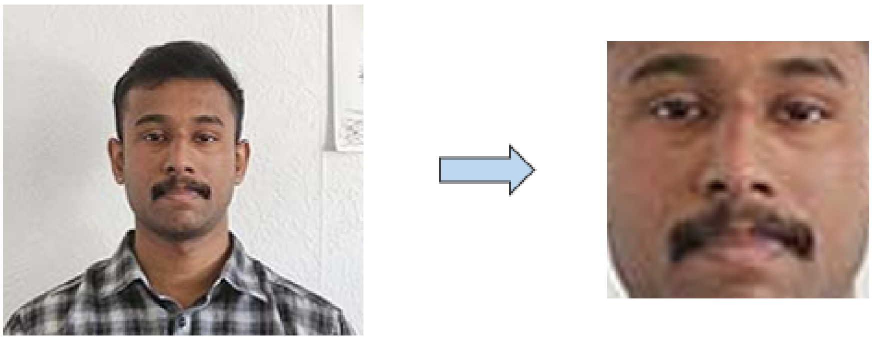


Figure 2.4: Example of preprocessing of image using FrontalFaceLBP, image before using FrontalFaceLBP(left), image after using FrontalFaceLBP (right).

2.5.3 Image resizing

Before inputting the facial images into network models such as ResNet, GoogLeNet, and AlexNet, the images are cropped in a manner that ensures the input size of each image aligns with the input size expected by the respective model, as represented in table 2.1.

Network Model	Image Size (pixels)
ResNet	224x224
GoogLeNet	224x224
AlexNet	227x227

Table 2.1: Image sizes for different network models

2.5.4 Adding noise to image dataset

Image noise refers to alterations in the image signal caused by external disturbances. In digital images, impulse noise, commonly known as Salt and Pepper noise, often occurs due to transmission errors [1]. The images with salt and pepper noise consists of infrequent occurrences of white and black pixels [4]. To evaluate the effect of dataset size variation, salt-and-pepper noise was introduced to 20% of the total pixels in an image within the processed dataset, effectively doubling its initial size. Figure 2.5 illustrates the impact on the image after adding the noise.

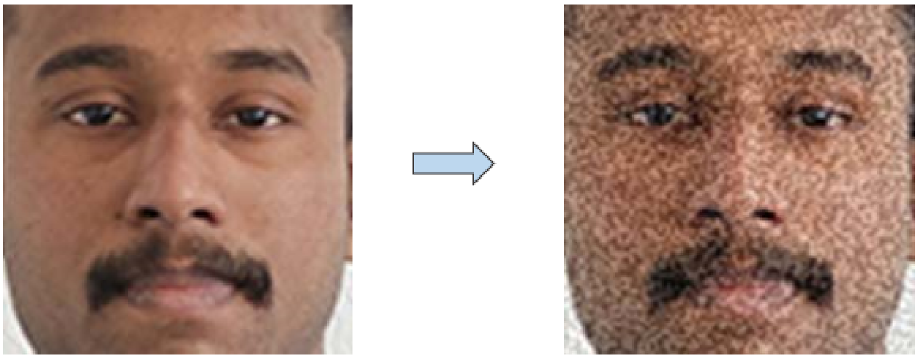


Figure 2.5: Effect of salt-and-Pepper Noise, image without noise (left), image after adding noise (right).

2.5.5 Adding brightness to image dataset

Recognition performance varies significantly across datasets with varying image qualities [21]. As some of the images of Oulu-CASIA [40] dataset exhibit artifacts, including low brightness, identifying them from a large dataset becomes challenging. To address this, the input database was expanded by enhancing the brightness (figure 2.6) of original images by a factor of 1.5.

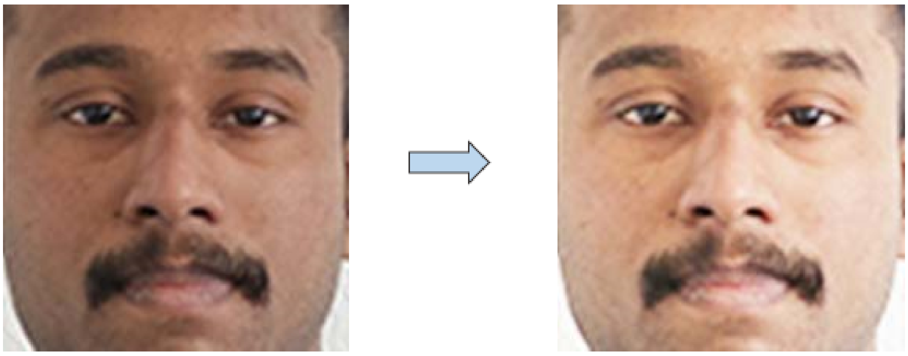


Figure 2.6: Adding brightness to the images, image before adding brightness (left), image after adding brightness (right).

2.6 Network model architecture design

Two different frameworks, named Model 1 and Model 2, have been constructed for the purpose of evaluating the performance of systems that recognize facial expressions.

2.6.1 Model 1

This model structure used three distinguished CNN designs such as GoogleNet, AlexNet, and ResNet50. The three architectures are explained as following.

Model architecture based on GoogleNet

The GoogleNet architecture incorporates a 22-layer deep network known as the Inception module as shown in figure 2.7 [15]. This module is specifically designed

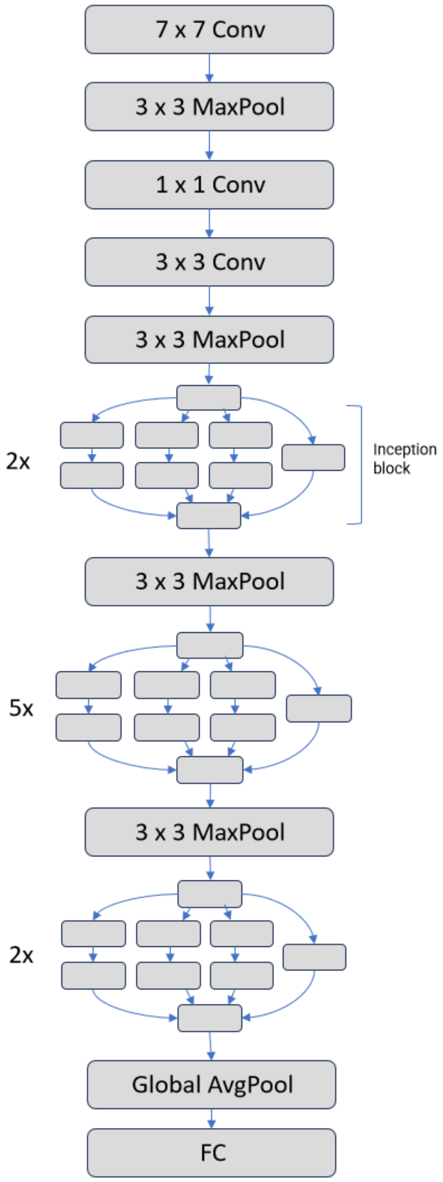


Figure 2.7: GoogleNet architecture.

to identify intricate patterns within input data. In this facial expression recognition model, input of images containing facial expressions and the corresponding labels with represented emotions are undergo processing by the GoogleNet model. The model classifies these images into seven emotion classes including "Neutral" emotion, and the final output is obtained from the *loss3-classifier* layer.

Model architecture based on ResNet50

ResNet50 is a powerful architecture used to address challenges in facial recognition. It consist of residual connections to allow gradients to flow directly through layers

[11] represented in figure 2.8. With 50 layers, including residual blocks, ResNet50 captures fine-grained facial features such as expressions, eye movements, and lip shapes [11]. In this study pre-trained ResNet50 models is also used for transfer learning, adapting them to specific facial recognition tasks using labeled Oulu-CASIA [40] dataset. The fully connected layer *fc1000* was modified and reduced to seven class representing each emotions. The features of image sequence is extracted from the *avg_pool* layer of ResNet50 architecture for training of the Model 2.

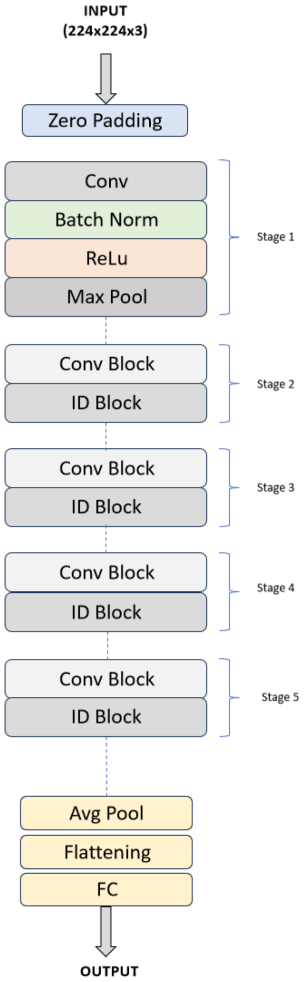


Figure 2.8: ResNet50 architecture.

Model architecture based on AlexNet

AlexNet consists of 25 layers with learnable parameters, and its input is RGB images. It composed of three convolutional layers, three fully connected layers, ReLU activations, and dropout layers [20]. The block representation of AlexNet is represented in figure 2.9. In this FER approch, AlexNet processes input images containing facial expressions, learn relevant features from the input images and classify output as seven emotions categorically. The *fc8* layer was converges to seven labels, similar to the previous approach used in both GoogleNet and ResNet50. for the purpose of feature extraction method *relu7* layer was considered.

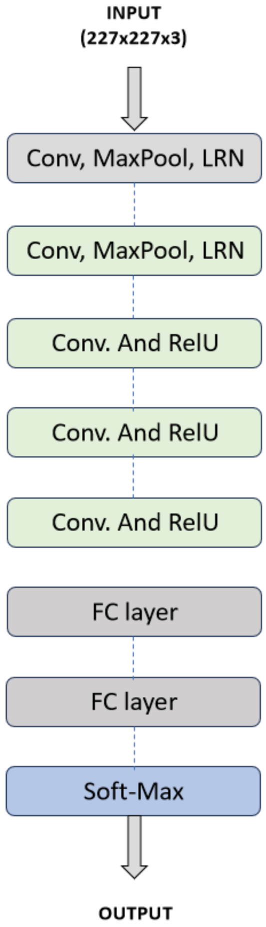


Figure 2.9: AlexNet architecture.

2.6.2 Model 2

Model 2 is created by utilizing features extracted from Model 1 after its training. This Model incorporates LSTM layers as in figure 2.10, which used features obtained from the last global average pooling layer of the corresponding network architecture used in Model 1. The features of the emotion sequence from AlexNet are obtained from the *relu7* layer. Similarly, the features from GoogLeNet and ResNet50 are extracted from the *loss3-classifier* layer and the *fc1000* layer, respectively. The primary goal of Model 2 is to evaluate the impact of LSTM layers on FER systems. By capturing temporal dependencies in image sequences, Model 2 aims to improve recognition accuracy. Assess and compare the performance of Model 2 with that of Model 1 using the Oulu-CASIA dataset.

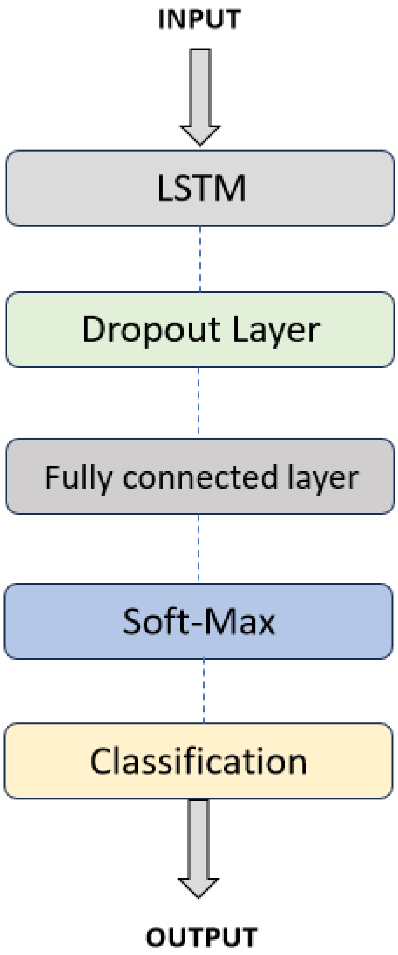


Figure 2.10: LSTM architecture.

2.7 Database

In this conducted study, the model was trained utilizing the Oulu-CASIA [40] image frame database. This database consists of images gathered from 80 participants, whose ages varied from 23 to 58 years. The individuals participated in recordings that showcased six different emotions such as, disgust, anger, fear, happiness, surprise, and sadness [40].

The captured emotional shifts in the database illustrate the transition from a neutral base to an overtly expressed emotion. At the beginning of each sequence, the first three frames were classified as neutral, leading to the identification of seven emotional categories for analysis. This methodology enabled a deeper examination of the emotional shifts and laid a solid basis for training the model.

For this study, images were carefully filtered to prioritize clarity and strong lighting conditions, resulting in a dataset consisting of 10,379 images [40]. Furthermore, the database is divided into three separate categories to assess the models' effectiveness. The category one consists of images from the primary dataset, categorized by emotional expression. The category two contains images from the primary dataset, alongside images that have been augmented with noise. The final category, includes images from the primary dataset, images augmented with noise, and images that have undergone an increase in brightness. The arrangement of these categories of images is illustrated in Figure 2.11.

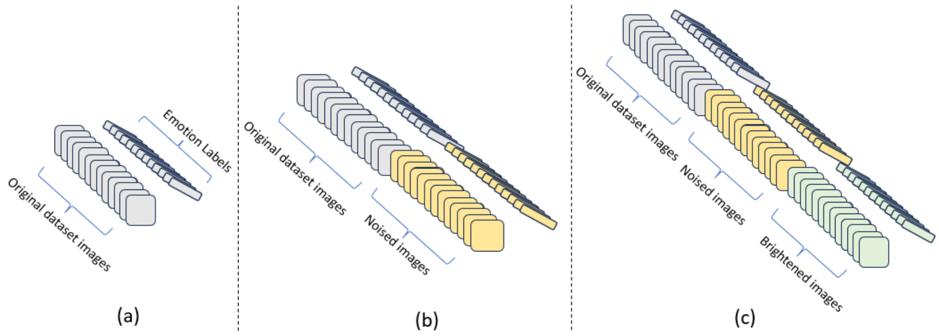


Figure 2.11: Arrangement of three different input categories of the database: (a) the first category with original images, (b) the second category with original and noised images, (c) the third category with original, noised, and brightened images.

2.8 Training process

2.8.1 Data allocation

The dataset used was divided into two subsets, training set and validating set. 70% of the data was allocated for training the models. This training set is used to adjust the model's parameters and optimize its predictive performance. The remaining 30% of the data was set aside as a validation set. This subset is used to evaluate

Table 2.2: Model 1 network training options

Parameter	Method / Value Selected
Solver	adaptive moment estimation (ADAM)
Mini Batch Size	50
# Epochs	40
Initial Learning Rate	0.0001

the model's performance during and after the training process.

2.8.2 Training of model 1

Model 1 is trained utilizing pre-processed images as its input. The training procedure is executed independently for each of the three categories of image sets, as specified in the dataset section. Moreover, the training incorporates three separate pre-trained neural networks, namely AlexNet, GoogleNet, and ResNet50, each network was trained by all three categories of the image set. The use of these preexisting networks offers a significant advantage as it substantially reduces the time required for training the model [6].

The training parameters employed for Model 1 are detailed in Table 2.2. The application of adaptive moment estimation (Adam) as a solver enables the model to manage larger data volumes while minimizing memory requirements [17]. The minibatch size of 50 and the number of epochs, set at 40, were determined based on resource constraints and preliminary training trials.

2.8.3 Training of model 2

Model 2 employs a distinct approach. Rather than using images directly, it used features derived from sequences of emotions. These features are derived from a particular layer in the model's architecture, situated just prior to the global average pooling layer of the networks trained in Model 1. This layers used for feature extraction of each architecture are different in size.

Table 2.3 outlines the training parameters for the development of Model 2. Similar to Model 1, the epoch and mini-batch size of 40 and 50, respectively, were determined based on resource availability and preliminary training trials in this study. The

Table 2.3: Model 2 network training options

Parameter	Method / Value Selected
Solver	stochastic gradient descent with momentum (SGDM)
Mini Batch Size	50
# Epochs	40
Initial Learning Rate	0.0001

optimization of Model 2 was achieved using the stochastic gradient descent with momentum (SGDM) algorithm. This algorithm is particularly effective in handling noise as it does not depend on the magnitude of gradient moments [10].

Chapter 3

Results

The effectiveness of model 1, was evalvated with networks - AlexNet, GoogleNet, and ResNet50, by employing three distinct batches of input images. This comparative analysis allowed for a comprehensive understanding of each model's performance under different conditions.

In addition to this, a specific focus was placed on Model 2. The objective was to examine how fluctuations in the number of output nodes influenced the model's ability to accurately recognize patterns or features. This involved a systematic alteration of the output nodes and monitoring the subsequent impact on the model's recognition accuracy.

3.1 Model 1

This model is subjected to a training process using a variety of image sets. These sets comprise the original image, a combination of the original and images with added noise, and a set with original, noised, and brightened images. To minimize the duration of the training phase, the model utilizes pretrained networks, namely GoogleNet, AlexNet, and ResNet50. Each network is trained with all three image sets.

3.1.1 Network performance with different sets of input data

In the experimentation phase, it was noted that training GoogleNet with the original image dataset yielded a recognition accuracy of 83.51%. However, incorporating both noisy and original images into the training dataset led to an improvement, resulting in a peak accuracy of 87.50%. This enhancement suggests the effectiveness of augmenting the training data with noisy samples alongside the original ones shown in figure 3.1. Comparable performance to GoogleNet was observed with ResNet50.

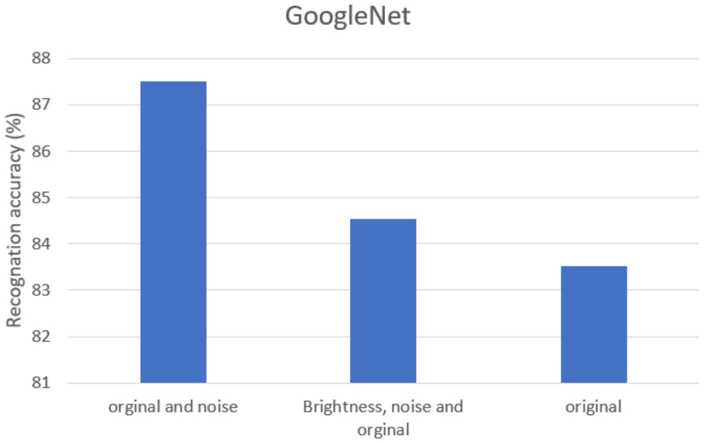


Figure 3.1: GoogleNet performance with different sets of input data.

The model achieved a maximum recognition accuracy of 84.48% when trained on input data comprising both original and noisy images (figure 3.2). Furthermore, when incorporating a combination dataset containing original, noisy, and brightened images, the model's performance improved to 84.19%, up from the baseline accuracy of 83.87% obtained with the original input image set.

AlexNet exhibited distinct performance characteristics compared to GoogleNet and ResNet50. When trained with a dataset containing original, noisy, and brightened images, AlexNet achieved its maximum recognition accuracy of 86.13%. Notably, the model's performance was lower when trained solely on original images, with a minimum accuracy of 85.20%. However, by augmenting the dataset with noisy images, the accuracy improved from 85.20% to 86.02%. This suggests that AlexNet benefits from the inclusion of both noise and brightness variations in the training data, leading to improved recognition accuracy. The performance data of AlexNet is shown in figure 3.3.

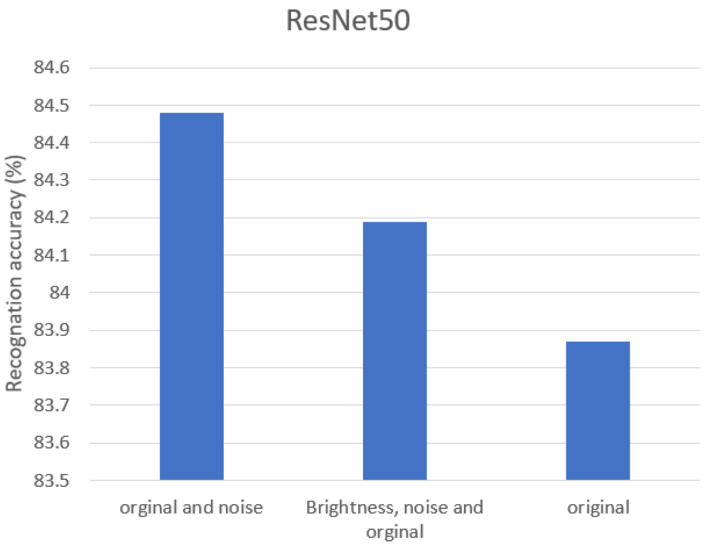


Figure 3.2: ResNet performance with different sets of input data.

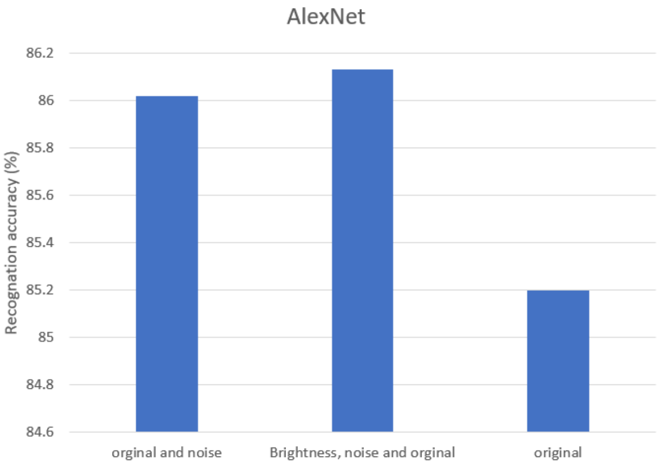


Figure 3.3: AlexNet performance with different sets of input data.

3.2 Model 2

Model 2 was trained by extracting features from each emotion sequence acquired from the layer preceding the final pooling layer of each network architecture. Testing of Model 2 was carried out on all three input image sets, along with variations of the model trained with different numbers of nodes in the LSTM layer, specifically 64, 128, and 250 nodes.

3.2.1 performance of GoogleNet with LSTM

From the figure 3.4, it is evident that the architecture of Model 2 enhanced the performance compared to the model trained with pretrained GoogleNet (referred to as Model 1). Model 1 achieved a minimum accuracy of 83.51%, which significantly improved to 94.21% when using Model 2 trained with 250 nodes.The LSTM architecture consistently achieves recognition accuracies ranging from 93.49% to 94.78%, regardless of the pre-trained models utilized.

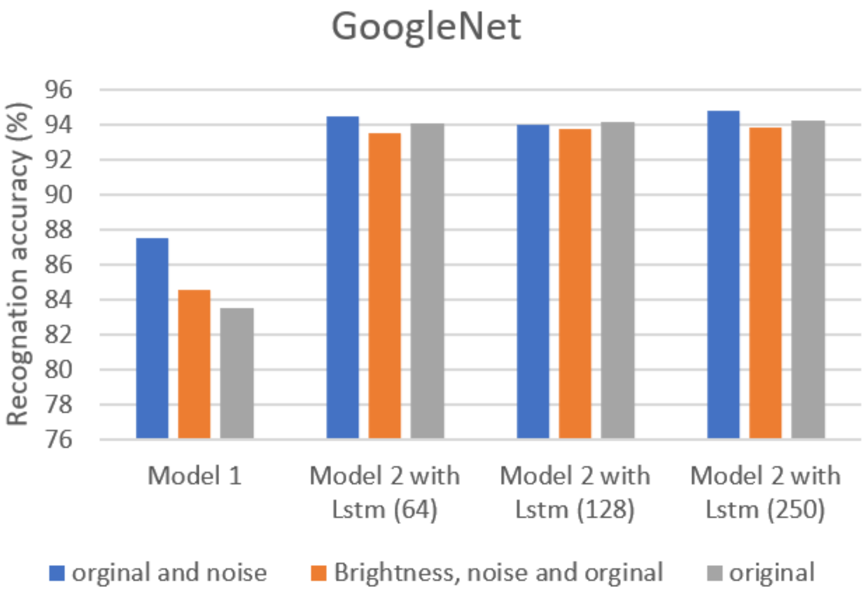


Figure 3.4: GoogleNet performance comparison

3.2.2 performance of AlexNet with LSTM

Model 1, employing AlexNet, attained a maximum accuracy of 86.13%. This accuracy was enhanced to a maximum of 93.63% by incorporating 250 nodes in LSTM. As depicted in Figure 3.5, it is evident that the number of nodes has a minor impact on enhancing the model's accuracy for emotion prediction. The LSTM architecture further boosted the accuracy of this model to a maximum of 94.06% when trained on original and noised images with 250 nodes.

3.2.3 performance of ResNet50 with LSTM

In comparison, LSTM outperformed GoogleNet and AlexNet when paired with ResNet50, achieving a maximum accuracy range of 94.99% to 96.40%. There were

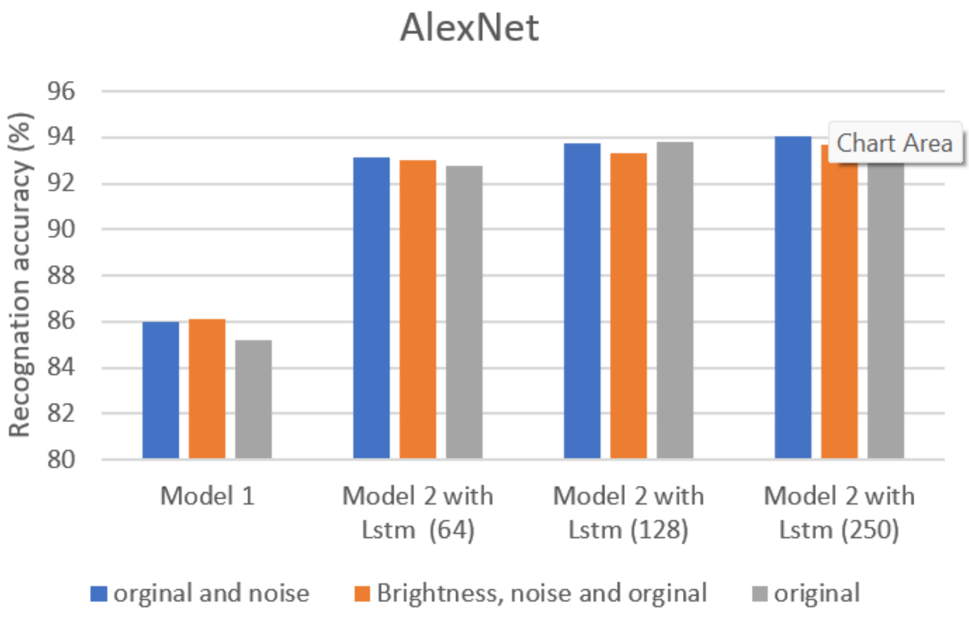


Figure 3.5: AlexNet performance comparison

no significant improvements observed with changes in the number of LSTM nodes; in fact, performance slightly decreased with increased node count. The peak accuracy of 96.40% was attained when utilizing a dataset comprising brightened, noised, and normal images with 64 nodes in LSTM. Figure 3.6 depicts the recognition accuracy across various input types and networks.

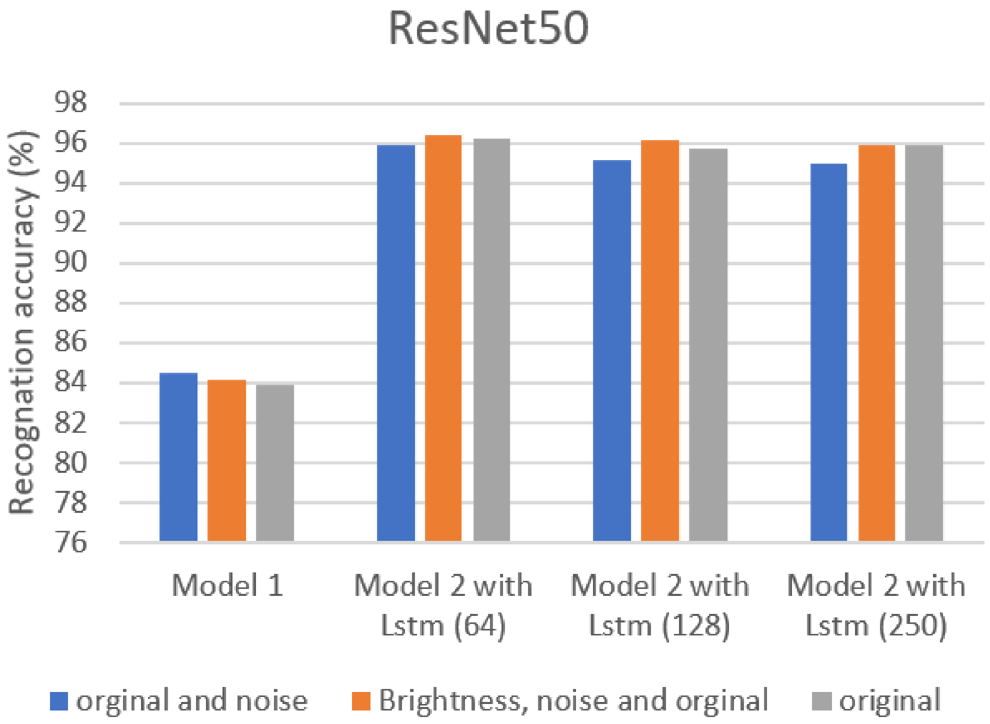


Figure 3.6: AlexNet performance comparison

Chapter 4

Discussion

The experimental outcomes provide significant understanding of the performance of deep learning architectures, specifically GoogleNet, ResNet50, and AlexNet, in the domain of FER. Furthermore, the investigation of Model 2, which combines LSTM layers with pretrained models, illuminates the influence of feature extraction and sequence modeling on the precision of FER.

4.0.1 Performance of Model 1

The assessment of Model 1 underscores the importance of data augmentation in enhancing recognition precision across various deep learning models. Data augmentation, a common method to artificially expand the variety of the training dataset, aids the model in better generalizing to unfamiliar data. When trained with a combination of original and noisy images, both GoogleNet and ResNet50 showed significant accuracy improvements compared to training exclusively on original images. This observation is consistent with prior studies that highlight the advantages of augmenting training data to increase model generalization and robustness. By introducing the model to variations in image quality and noise levels, data augmentation enables the model to learn more resilient features, thereby increasing its adaptability to environmental changes that may occur during real-world application.

In contrast with, AlexNet's performance exhibited a distinct pattern, reaching its peak accuracy when trained on a dataset comprising original, noisy, and brightened images. This implies that AlexNet might be more tolerant of variations in image quality and lighting compared to GoogleNet and ResNet50. The noted accuracy increase upon augmenting the dataset with noisy images further emphasizes the significance of including diverse samples to enhance model performance. These findings highlight the importance of selecting appropriate data augmentation strategies tailored to the characteristics of the model architecture and the nature of the task.

4.0.2 Performance of Model 2

The incorporation of LSTM layers in Model 2 significantly improves recognition accuracy across all pretrained models. LSTM, a form of RNN, is adept at processing sequential data, such as time series analysis and natural language processing. In the field of FER, LSTM allows the model to access temporal dependencies in facial emotion sequences, thereby effectively learning the progression of emotions over time. This enhancement is particularly evident when using ResNet50, where LSTM consistently surpasses the baseline accuracy achieved by Model 1. The relatively consistent performance of LSTM across varying node numbers suggests that the LSTM architecture capably captures temporal dependencies in facial emotion sequences, irrespective of network complexity.

The comparison of LSTM performance across pretrained networks reveals intriguing insights into their suitability for sequence modeling in FER. While LSTM increases the accuracy of all models, ResNet50 displays the best performance, attaining a peak accuracy of 96.40%. This superior performance could be attributed to ResNet50's deeper architecture and sophisticated feature extraction abilities, which offer high representations for LSTM to effectively learn temporal patterns. These observations emphasize the importance of utilizing both spatial and temporal information in FER tasks, underscoring the synergistic relationship between deep learning architectures and sequence modeling techniques.

Chapter 5

Conclusions

In summary, this research explores the use of deep learning methods for FER specifically for individuals with ASD. The study provides insightful findings on the efficacy of data augmentation, pretrained models, and sequence modeling in enhancing FER precision and real-time functionality for this particular group.

The outcomes highlight the importance of data augmentation strategies, such as the inclusion of noisy and brightened images, in boosting FER precision across various deep learning models, including GoogleNet, ResNet50, and AlexNet. By subjecting the models to variations in image quality and lighting, data augmentation contributes to improved model generalization and robustness, which is especially crucial for individuals with ASD who may display a wide range of facial expressions.

Moreover, the fusion of Long Short-Term Memory (LSTM) layers with pretrained networks notably improves recognition precision, especially when using ResNet50. LSTM allows the model to identify temporal dependencies in facial emotion sequences, thereby effectively learning the progression of emotions over time. This enhancement emphasizes the importance of utilizing both spatial and temporal data in FER tasks for individuals with ASD.

The result of this study indicate that Model 2, integrating LSTM layers with ResNet50, emerges as the most suitable model for detecting emotions from facial images. The incorporation of LSTM facilitates the capture of sequential features of each emotion, enhanced the precision of recognition. ResNet50 consistently outperforms other pretrained networks such as AlexNet and GoogleNet, underscoring its efficacy in extracting facial features essential for accurate emotion recognition. Moreover, in real-time testing circumstances, Model 2 appears to be a better model due to its high recognition accuracy, which makes it a good fit for practical applications aimed at assisting people with ASD.

Overall, this research contributes to the advancement of understanding and application of deep learning methods for FER in individuals with ASD. By employing data augmentation, pretrained models, and sequence modeling, the developed models present promising pathways for the creation of intelligent systems capable of comprehending and responding to the unique emotional expressions of individuals with ASD. However, additional research is needed to tackle challenges such as interpretability, ethical considerations, and real-world implementation, to ensure the responsible and effective utilization of FER technology in assisting individuals with ASD and fostering their social and emotional growth.

Future work

Future research in the context of individuals with Autism Spectrum Disorder (ASD) could explore advanced data augmentation strategies tailored to the unique characteristics of facial expressions exhibited by individuals with ASD. For instance, employing generative adversarial networks (GANs) could enable the generation of realistic variations in facial expressions specific to ASD, considering the distinct patterns in their emotional expressions. By incorporating ASD-specific facial expression data into the training process, models can be better adapted to recognize and interpret the subtle cues indicative of different emotions in individuals with ASD, thereby enhancing model generalization and performance.

Moreover, the integration of physiological data sources, such as Electrocardiography (ECG) and Electromyography (EMG), holds promise for enriching the contextual information available for emotion recognition in individuals with ASD. Physiological signals can provide additional insights into the internal states and arousal levels associated with specific emotions, complementing the facial expression data and contributing to a more comprehensive understanding of emotional responses in indi-

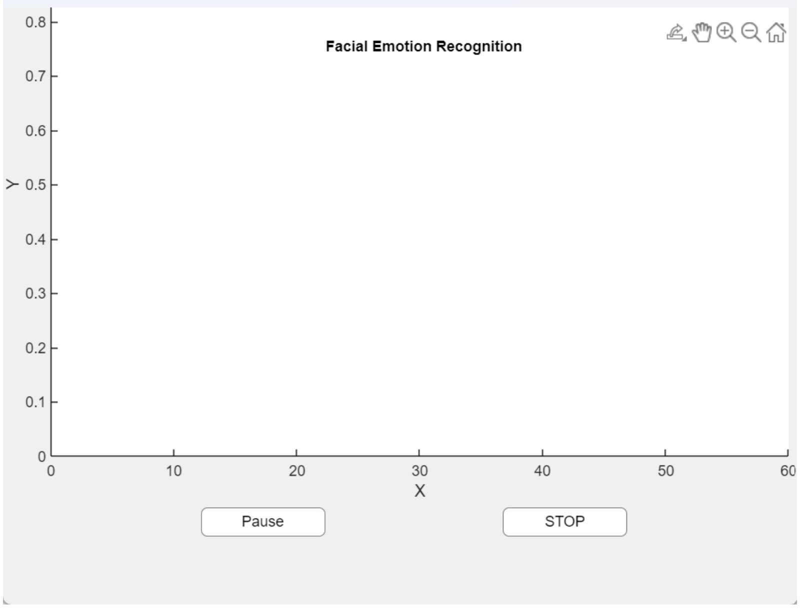
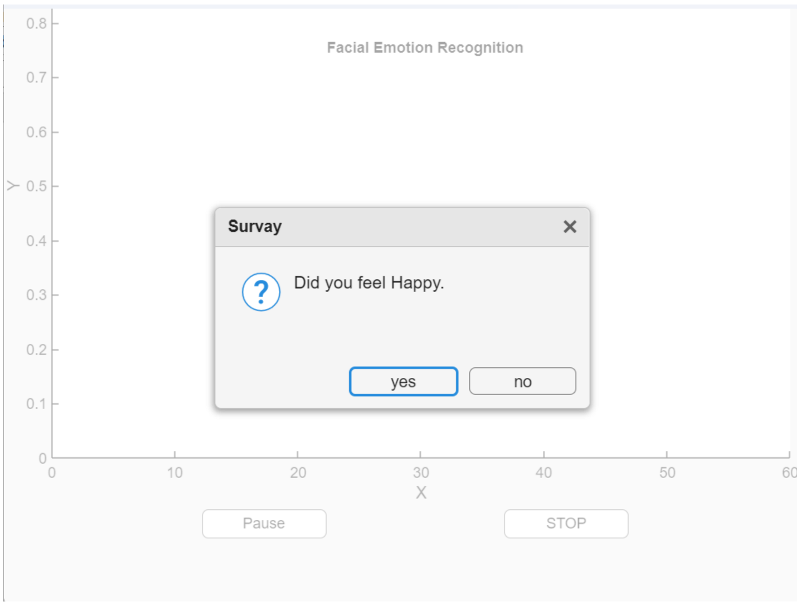


Figure 5.1: Proposed App view during emotion recording.

viduals with ASD. By leveraging multimodal data fusion techniques, future models can benefit from the synergistic combination of facial expressions and physiological signals, leading to more accurate and robust emotion recognition models tailored to the needs of individuals with ASD.

Furthermore, live testing serves as an essential step for further validating the models, ensuring their performance extends beyond controlled environments to real-world scenarios. A proposed framework for conducting live testing is illustrated in Figures 5.1 and 5.2. This live testing initiative is vital for gaining valuable insights into the practical applicability and efficacy of the models. Such insights will be instrumental in guiding future enhancements and refinements to optimize the models for real-world deployment.



Survay

Did you feel Happy.

Figure 5.2: Proposed App view during validation of emotion by subject.

References

[1] Naif Alajlan, Mohamed Kamel, and Ed Jernigan. "Detail preserving impulsive noise removal". In: *Signal processing: image communication* 19.10 (2004), pp. 993–1003.

[2] Laith Alzubaidi et al. "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions". In: *Journal of big Data* 8 (2021), pp. 1–74.

[3] Herag Arabian et al. "Image Pre-processing Significance on Regions of Impact in a Trained Network for Facial Emotion Recognition". In: *IFAC-PapersOnLine* 54.15 (2021), pp. 299–303.

[4] Jamil Azzeh, Bilal Zahran, and Ziad Alqadi. "Salt and pepper noise: Effects and removal". In: *JOIV: International Journal on Informatics Visualization* 2.4 (2018), pp. 252–256.

[5] Baeldung. *Neural Network Pre-training.* `https://www.baeldung.com/cs/neural-network-pre-training`. Accessed: 2024-02-24.

[6] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. "Net2net: Accelerating learning via knowledge transfer". In: *arXiv preprint arXiv:1511.05641* (2015).

[7] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. "Deep learning-based facial emotion recognition for human–computer interaction applications". In: *Neural Computing and Applications* 35.32 (2023), pp. 23311–23328.

[8] Akhmedov Farkhod et al. "Development of real-time landmark-based emotion recognition CNN for masked faces". In: *Sensors* 22.22 (2022), p. 8704.

[9] Beat Fasel and Juergen Luettin. "Automatic facial expression analysis: a survey". In: *Pattern recognition* 36.1 (2003), pp. 259–275.

[10] Aman Gupta et al. “Adam vs. sgd: Closing the generalization gap on image classification”. In: *OPT2021: 13th Annual Workshop on Optimization for Machine Learning*. 2021.

[11] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[12] Shantala S Hiremath et al. “Facial Expression Recognition Using Transfer Learning with ResNet50”. In: *Inventive Systems and Control: Proceedings of ICISC 2023*. Springer, 2023, pp. 281–300.

[13] IBM. *Convolutional Neural Networks*. Accessed: 2023-10-26. 2024. URL: `https://www.ibm.com/topics/convolutional-neural-networks`.

[14] Zifan Jiang et al. “Classifying major depressive disorder and response to deep brain stimulation over time by analyzing facial expressions”. In: *IEEE transactions on biomedical engineering* 68.2 (2020), pp. 664–672.

[15] P Kalaiarasi and P Esther Rani. “A comparative analysis of AlexNet and GoogLeNet with a simple DCNN for face recognition”. In: *Advances in Smart System Technologies: Select Proceedings of ICFSST 2019*. Springer. 2021, pp. 655–668.

[16] Yousif Khaireddin and Zhuofa Chen. “Facial emotion recognition: State of the art performance on FER2013”. In: *arXiv preprint arXiv:2105.03588* (2021).

[17] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[18] Byoung Chul Ko. “A brief review of facial emotion recognition based on visual information”. In: *sensors* 18.2 (2018), p. 401.

[19] Alex Krizhevsky. “One weird trick for parallelizing convolutional neural networks”. In: *CoRR* abs/1404.5997 (2014). arXiv: `1404.5997`. URL: `http://arxiv.org/abs/1404.5997`.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60 ().

[21] Ke Li et al. “Sharpness and brightness quality assessment of face images for recognition”. In: *Scientific Programming* 2021 (2021), pp. 1–21.

[22] Shanshan Li, Liang Guo, and Jianya Liu. “Towards East Asian facial expression recognition in the real world: A new database and deep recognition baseline”. In: *Sensors* 22.21 (2022), p. 8089.

[23] Yanli Liu, Yuan Gao, and Wotao Yin. “An improved analysis of stochastic gradient descent with momentum”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 18261–18271.

[24] Catherine Lord et al. “Autism spectrum disorders”. In: *Neuron* 28.2 (2000), pp. 355–363.

[25] Marius Mosbach et al. “Few-shot Fine-tuning vs. In-context Learning: A Fair Comparison and Evaluation”. In: *arXiv preprint arXiv:2305.16938* (2023).

[26] T. Ojala, M. Pietikainen, and T. Maenpaa. “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7 (2002), pp. 971–987. DOI: 10.1109/TPAMI.2002.1017623.

[27] OpenCV. *Cascade Classifier*. Accessed: 2024-02-26. 2024. URL: https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html.

[28] G Padmashree and AK Karunakar. “Improved LBP Face Recognition Using Image Processing Techniques”. In: *Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces*. Springer, 2022, pp. 535–546.

[29] Seo-Jeon Park, Byung-Gyu Kim, and Naveen Chilamkurti. “A robust facial expression recognition algorithm based on multi-rate feature fusion scheme”. In: *Sensors* 21.21 (2021), p. 6954.

[30] Alexander M Pascual et al. “Light-FER: A Lightweight Facial Emotion Recognition System on Edge Devices”. In: *Sensors* 22.23 (2022), p. 9524.

[31] Muhammad Salihin Saealal et al. “Using cascade CNN-LSTM-FCNs to identify AI-altered video based on eye state sequence”. In: *PLoS One* 17.12 (2022), e0278989.

[32] Hasim Sak, Andrew W Senior, and Françoise Beaufays. “Long short-term memory recurrent neural network architectures for large scale acoustic modeling”. In: (2014).

[33] Rajesh Singh et al. "Facial expression recognition in videos using hybrid CNN & ConvLSTM". In: *International Journal of Information Technology* 15.4 (2023), pp. 1819–1830.

[34] Xiao Sun, Pingping Xia, and Fuji Ren. "Multi-attention based deep neural network with hybrid features for dynamic sequential facial expression recognition". In: *Neurocomputing* 444 (2021), pp. 378–389.

[35] Christian Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015, pp. 1–9.

[36] Yuta Takahashi et al. "Neural network modeling of altered facial expression recognition in autism spectrum disorders based on predictive processing framework". In: *Scientific reports* 11.1 (2021), p. 14684.

[37] Monu Verma et al. "Automer: Spatiotemporal neural architecture search for microexpression recognition". In: *IEEE Transactions on Neural Networks and Learning Systems* 33.11 (2021), pp. 6116–6128.

[38] Paul Viola and Michael Jones. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001.* Vol. 1. Ieee. 2001, pp. I–I.

[39] Kaihao Zhang et al. "Facial expression recognition based on deep evolutional spatial-temporal networks". In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4193–4203.

[40] Guoying Zhao et al. "Facial expression recognition from near-infrared videos". In: *Image and vision computing* 29.9 (2011), pp. 607–619.

[41] Xiaoming Zhao and Shiqing Zhang. "A Review on Facial Expression Recognition: Feature Extraction and Classification". In: *IETE Technical Review* 33.5 (2016), pp. 505–517. DOI: `10.1080/02564602.2015.1117403`.

[42] Yufeng Zheng and Erik Blasch. "Facial Micro-Expression Recognition Enhanced by Score Fusion and a Hybrid Model from Convolutional LSTM and Vision Transformer". In: *Sensors* 23.12 (2023), p. 5650.