AUDIO-VISUAL EMOTION RECOGNITION USING DEEP OPERATIONAL NETWORKS

DERİN OPERASYONEL AĞLAR İLE İŞİTSEL-GÖRSEL DUYGU TANIMA

KAAN AKTÜRK

ASSOC. PROF. ALI SEYDI KEÇELİ

Supervisor

Submitted to

Graduate School of Science and Engineering of Hacettepe University

as a Partial Fulfillment to the Requirements

for the Award of the Degree of Master of Science

in Computer Engineering

January 2024

ABSTRACT

AUDIO-VISUAL EMOTION RECOGNITION USING DEEP OPERATIONAL NETWORKS

Kaan AKTÜRK

Master of Science , Computer Engineering

Supervisor: Assoc. Prof. Ali Seydi KEÇELİ

January 2024, 74 pages

Emotions have a significant impact on interpersonal communication, marketing, healthcare, and the service sector. Consequently, much study continues to be conducted on the categorization of emotions up to the present day. Audio-visual emotion detection is a common area of study in the realm of machine learning. Its primary objective is to identify and categorize human emotions. It utilizes computer vision and audio processing methods to assess and understand the emotional states conveyed by people. But most of the studies are conducted by analyzing one type of data, such as texts or images. This research introduces an operational neural network-based deep learning model that utilizes various inputs to provide emotion identification. The proposed model employs a comprehensive strategy that incorporates both visual and audio features. The suggested architecture substitutes conventional convolutional layers with operational layers. The experimental findings indicate that the operational convolutional architecture outperforms the traditional convolutional neural network design.

Keywords: emotion classification, multi-input classification, operational neural network, audio-visual classification, visual geometry group

ÖZET

DERİN OPERASYONEL AĞLAR İLE İŞİTSEL-GÖRSEL DUYGU TANIMA

Kaan AKTÜRK

Yüksek Lisans, Bilgisayar Mühendisliği

Danışman: Assoc. Prof. Ali Seydi KEÇELİ

Aralık 2023, 74 sayfa

Duyguların kişilerarası iletişim, pazarlama, sağlık hizmetleri ve hizmet sektörü üzerinde önemli bir etkisi vardır. Bu nedenle duyguların sınıflandırılması konusunda günümüze kadar pek çok çalışma yapılmaya devam etmektedir. İşitsel-görsel duygu tespiti, makine öğrenimi alanında yaygın bir çalışma alanıdır. Temel amacı insan duygularını tanımlamak ve sınıflandırmaktır. Kişilerin aktardığı duygusal durumları değerlendirmek ve anlamak için bilgisayarda görü ve ses işleme yöntemlerinden yararlanır. Ancak çalışmaların çoğu metin ya da görsel gibi tek tip veriyi analiz ederek yürütülüyor. Bu araştırma, duygu tanımlamayı sağlamak için çeşitli girdileri kullanan operasyonel sinir ağı tabanlı bir derin öğrenme modelini tanıtmaktadır. Önerilen model, hem görsel hem de işitsel özellikleri birleştiren kapsamlı bir strateji kullanmaktadır. Önerilen mimari, geleneksel evrişim katmanlarını operasyonel katmanlarla değiştirmektedir. Deneysel bulgular, operasyonel evrişimli mimarinin, geleneksel evrişimli sinir ağı tasarımından daha iyi performans göstermektedir.

Keywords: duygu sınıflandırma, çok girdili duygu sınıflandırma, operasyonel nöral ağlar, işitsel-görsel sınıflandırma

ACKNOWLEDGEMENTS

I am profoundly grateful to Ali Seydi Keçeli for his unwavering support and for sharing his experiences throughout this process. Working with you has been a privilege for me.

I dedicate this work to my family for their endless support and love. I also want to thank them for tolerating my excuses of writing this thesis, while at the same time managing to overlook my avoidance of many family responsibilities. After all, becoming a high engineer is no easy feat.

I would also like to send my love to my best friends. Despite constantly judging all of you, I appreciate your patience in listening to me. In reality, since I carried all your problems on my shoulders, you should be thanking me. But anyway, we made it here together in the end.

Finally, to myself, Are you aware of the situation? You've reached the end of another goal. Could you have predicted this? Despite facing internal struggles that you kept to yourself, you managed to overcome this process. Another item is crossed off the list of difficulties. So, the main question is... As Taylor Swift once said, ***"If the story's over, why am I still writing pages? 'Cause saying goodbye is death by a thousand cuts."***

Thank you, Next!

Kaan Aktürk

CONTENTS

Page

TABLES

FIGURES

ABBREVIATIONS

BERT	:	**B**idirectional **E**ncoder **R**epresentations **T**ransformers
CNN	:	**C**onvolutional **N**eural **N**etwork
CMYK	:	**C**yan **M**agenta **Y**ellow **B**lack
FACS	:	**F**acial **A**ction **C**oding **S**ystem
LIF	:	**L**eaky **I**ntegrate **F**ire
LSTM	:	**L**ong **S**hort **T**erm **M**emory
MFCC	:	**M**el **F**requency **C**epstral **C**oefficients
MLP	:	**M**ulti **L**ayer **P**erceptron
NLP	:	**N**atural **L**anguage **P**rocessing
ONN	:	**O**perational **N**eural **N**etwork
RGB	:	**R**ed **G**reen **B**lue
SSL	:	**S**elf **S**upervised **L**earning
SNN	:	**S**pike **N**eural **N**etwork
STDP	:	**S**pike **T**ime **D**ependent **P**lasticity
SVM	:	**S**upport **V**ector **M**achines
TF	:	**T**erm **F**requency
GPT	:	**G**enerative **P**re-trained **T**ransformer
IDF	:	**I**nverse **D**ocument **F**requency
RNN	:	**R**ecurrent **N**eural **N**etwork
SPL	:	**S**ingle **L**ayer **P**erceptron
TF-IDF	:	**T**erm **F**requency-**I**nverse **D**ocument **F**requency
VSM	:	**V**ector **S**pace **M**odel
VGG	:	**V**isual **G**eomtry **G**roup

1. INTRODUCTION

1.1. Background of Emotion Classification

The classification of emotions is a method that classifies human feelings according to the context in which they are experienced. These states may comprise data conveyed by vocalizations, text communications, facial expressions, and body language cues. The objective of the classification is to determine an individual's emotional state under specific circumstances and at a particular time. One of the most pressing concerns for scholars today is how to categorize different feelings. It has been studied in a variety of disciplines, and it has the potential to deliver significant benefits. Emotion classification can be put to use in a variety of contexts, including marketing and brand monitoring, financial analysis, monitoring of social media, health care, and customer service. In this work, the importance of emotion categorization and its application areas are investigated, and a neural network model is presented as a potential solution to this issue. The basic purpose of emotion classification is to gain an understanding of the emotional state of the individual [1]. Some of the researchers divide the states of emotion into three categories: neutral, positive, and negative. Moreover, using more complex classification techniques, the emotions can also be classified with some emotion states such as angry, pleased, sad, and surprised.

1.2. Usage of Emotion Classification

The use of sentiment analysis significantly aids in the evaluation and comprehension of client feedback. It presents some opportunities and benefits for businesses. The classification of emotions makes it possible to distinguish between positive and negative statements in customer reviews. It contributes to a better understanding of the customer's feelings toward the product. Classification is one method via which businesses can submit feedback. It is also helpful to indicate which aspects of the products are liked by the customers who have left positive remarks. Therefore, it can assist businesses in recognizing their shortcomings and giving them the opportunity to make enhancements in the relevant areas [1]. Additionally,

it has the potential to make a number of important contributions to the field of medicine. The classification of emotions might have an impact on how diseases are diagnosed, how satisfied patients are with the services they receive, and how therapy is carried out. To begin, the evaluation of the treatment that a patient receives at a healthcare facility might incorporate the classification of emotions experienced by the patient. Moreover, evaluation and categorization of the user comments provided by the patient for the service are both possible. When patients' medical reports are analyzed, one can gain information on the patients' emotional and psychological states. Patients may benefit from receiving improved and more in-depth treatment through the use of this strategy [2]. A further use of the emotion classification can be found on many social media platforms [3]. Today, we can witness the effects of emotion classification on social media platforms, where there is a lot of data collection. On social media, users frequently discuss their personal lives and events. Every day, users produce millions of new records in social media databases. This newly available big data can assist in understanding societal trends through the classification of emotions and provide convenience when it comes to adopting safeguards. Additionally, analyzing the content that they share on social media and the comments that users leave on the advertisements that they publish allows companies to make a variety of assessments about their products. Evaluations might be made of analyses that are based on unfavorable feelings. It is possible for businesses to participate in product enhancement processes.

1.3. Challenges in Emotion Classification

The categorization of emotions based on facial expressions poses considerable issues due to the variety of aspects involved. These aspects could potentially have some negative consequences for deep learning models as well.

Humans are able to feel more than one emotion at the same time and reflect their experiences through their facial expressions instantly. For example, a person may frown at one moment and smile at the next moment. Consequently, the categorization of emotions may be called into doubt when many sentiments coexist.

Another challenge is that the same facial expression can have quite varied meanings depending on where you are in the world and what culture you are a part of [4]. It's possible for people from various parts of the world to express the same feeling in very diverse ways. As a consequence of this, there is a possibility that there will be problems in the emotion classification analyses done by people who are exposed to varied social situations. Additionally, the use of irony and comedy has a substantial impact on the interpretation of emotions. It's possible that the irony and humor will lead to some confusion [5]. It is possible for techniques of classifying emotions to produce errors when the individual attempting to communicate an emotion does not do so in a very clear and concise manner.

1.4. Objectives of the Thesis

The examination of facial expressions via images, as well as audio and written records, can be utilized in the process of emotion classification. However, due to the challenges experienced during emotion classification stated above, it is possible that conducting emotion analysis using only a single category of data, such as facial expressions or speech transcripts, may not be sufficient. In the process of analyzing people's emotions, not only can we analyze the facial expressions of individuals, but we can also evaluate the vocal signals that people create when they make expressions.

On the topic of emotion classification, some research has presented a few different ways. Some of their presented solutions are based on only text data, while others are based on data derived from audio or images. On the other hand, there is some research that presents multiple types of data that are taken into consideration as potential inputs to their model.

In this work, we investigate a technique for emotion classification known as the multi-input model. The classification model can be approached using the multi-input emotion classification method, which involves delivering more than one input at the same time. In many ways to classify emotions, merely the text or pictures are studied, and then predictions are formed based on those findings. In the current investigation, it is intended to make use of two different inputs for the model. The first input is a visual representation of the person's

facial expression. The second one is the visual of the individual's vocal pattern signal. We anticipate the outcomes to be more varied and the information to be more thorough when more than one input is delivered to the developed model. Furthermore, the suggested model utilizes neural network methodologies. However, this study employs non-linear computations as opposed to the linear calculations utilized in standard neural networks. Additionally, the study aims to investigate the impact of these non-linear calculations on the categorization procedure.

1.5. Organization

The study is constructed in the following manner. The following section digs into the scholarly research undertaken on emotion categorization in the academic literature. Both the techniques utilized in these investigations and the evaluations of the methodologies are offered. The third section of our analysis outlines the techniques employed. The approaches are thoroughly discussed. The subsequent section presents the generated model of this investigation. The methodology for classifying numerous emotions based on input is explained and examined. The study discusses the data used as inputs for the model and the neural network topologies employed by the model. The fifth section of the document provides a detailed discussion of the data sets that are utilized for training and testing the model. The utilized data sets are described. Multiple figures are provided to visually represent the data sets. Then, the designated data set is used to assess our proposed model, and the results are then given. Then, the study's findings are consolidated.

2. LITERATURE OVERVIEW

In this part of the study, some research that has been conducted on the categorization of emotions is discussed. Emotions play an essential part in human interactions and the processes by which decisions are made. Because of this, the categorization of emotions has become an important task in a variety of domains, including human-computer interaction and psychology. It is also a field that draws from a variety of disciplines, and its significance

has been growing and expanding over the past few years. On this topic, there have been a lot of studies conducted. The applied studies investigate a variety of strategies that are currently employed in emotion categorization, ranging from more conventional methods to today's deep learning techniques. In this part, the methodology, conclusions, and problems associated with emotion classification will be discussed.

The study by Paul Ekman is considered to be the origin of traditional techniques [6]. The Facial Action Coding System (FACS) is the primary foundation upon which the classification of emotions is built. The system analyzes human facial expressions to recognize and classify a range of feelings and states of mind. His approach utilizes observable facial clues in conjunction with the FACS to categorize feelings. He investigates 6 fundamental feelings, which are sadness, anger, happiness, fear, and surprise, as well as disgust. The FACS classification methods for emotions are as follows:

- A happy smile can be seen on a person's face when the cheeks and the corners of the lips are pulled upward.
- The facial expression of sadness is characterized by a downward-turned mouth and an inner eyebrow that is dragged upward,.
- The facial expression of anger is characterized by lowered brows, narrowed eyes, and tightened lips.
- The expression of fear is characterized by eyes that are wider open, eyebrows that are elevated, and an open mouth.
- The expression of surprise is characterized by wide-open eyes, an open mouth, and eyebrows that are raised.
- The expression of disgust is characterized by a wrinkled nose and a lifted upper lip.

The premise of the research is that each of these feelings is associated with a unique facial cue that can be recognized by everyone. In addition to this, it is important to keep in mind that

these facial cues can hide more complicated emotional states. The research has a significant bearing on anthropology, psychology, and nonverbal communication, among other subjects. It also demonstrates how various cultural norms influence the language used.

Another study provides an explanation of the concept of extracting and assessing comments and feelings from text data [7]. It discusses the growing significance of opinion mining in the context of analyzing user feedback in a variety of applications, including social media. In the research, the challenges that are encountered during the analysis are discussed. In light of these challenges, it is clear how vital it is to give careful attention to the particulars of the language one uses and to select words that convey emotion effectively. In the context of supervised learning strategies, he investigates Support Vector Machines (SVM) as well as Naive Bayes methods. It has been stated that using labeled data sets is essential for classification in order to achieve more accurate findings. The terms ”positive” and ”negative” are used to describe feelings rather than descriptors like ”happy,” ”angry,” and so on. In unsupervised learning approaches, studies are carried out on emotion polarity using emotion words and normal words. The presented research offers a comprehensive summary of the strategies that have been utilized and the difficulties that have been encountered in the subject of emotion analysis. For research in the field of natural language processing (NLP), it has been demonstrated to be a valuable resource.

Expressions found within text data can also be analyzed to determine an individual's emotional state. In relation to this topic, the research literature contains many papers. In this study, common approaches for emotion categorization using text data are presented [8]. Within the field of NLP, for instance, Word Embedding is a significant research area. It displays words as continuous vectors in a multidimensional space and then converts the words into numerical form so that they may be used as input for machine learning models. The vectors establish syntactic and semantic connections between the words. After that, they use NLP models to determine the meaning of words by placing them in the appropriate context. Word2Vec and GloVe are two examples of algorithms that are discussed in the paper. They are produced using a large text corpus as a resource. So that they might compose sets of phrases that have a comparable meaning to one another.

In addition to that, the term frequency-inverse document frequency is discussed. The Term Frequency-Inverse Document, commonly known as TF-IDF, is another method that can be utilized to analyze an individual's feelings. It is able to determine the significance of a word in a given text. It has two units. The first is the term frequency (TF), and the second is the inverse document frequency (IDF). They are used to refer to the weights of the various words throughout the text.

Methods of machine learning can also be utilized in order to investigate the feelings conveyed by a document. As an illustration, the Vector Space Model (VSM) is an approach to natural language processing that investigates text data. It transforms the information contained in the text document into numerical vectors. The documents are presented in vector form. The words are presented in dimensional form. A matrix is organized in a structure that is formed from vectors and dimensions. Rows are used to present a document, whereas columns are used to shape individual words. When comparing the similarities between the sentences, the cosine similarity measure is utilized. Encoding the phrases in a sentiment vectoral space allows for the determination of the feelings contained within them.

It is also possible to assess pictures in order to categorize feelings. There are various models that can be constructed using body language and facial emotions. There are also some articles that analyze facial images in order to determine the emotions being conveyed. As an illustration, a model that is constructed using a Convolutional Neural Network (CNN) is shown [9]. In the discipline of deep neural networks, CNN is a study field that is commonly utilized for the analysis of images and videos. It includes numerous layers and is designed to learn the patterns and attributes of images. Different functions are served by each successive layer. It begins by retrieving features from the image and then applying filters such as edge detection, blurring, and sharpening. Following the applied filters to the image, patterns from the images are gathered. The connected layers use the information they've acquired to make predictions. An emotion categorization task is carried out by a CNN model in this article [9]. 4 different types of emotions which are disgusted, surprised, happy, and sad are used. The overall accuracy of the model is 92.50% on average.

W. Ragheb and colleagues have created a model for the analysis of sentiment that makes use of 3 distinct types of emotional expression [10]. The analyses are carried out on the model's text-based data. In terms of F1-score, it has achieved success at a rate of 75%. In addition, V. Kranthi Sai Reddy has carried out research on sentiment analysis using audio data [11]. The model of a multi-layered neural network is presented here. Using the AdaBoost ML approach, the signals extracted from the speech of individuals belonging to a particular group are categorized. The result has an accuracy of roughly 93%. Swayam Badhe et al. has also conducted research on a model that consists of four CNN layers and 2 fully connected levels [12]. This model is created using photos that contained different facial expressions. It has an accuracy rating of roughly 94 percent. Moreover, some studies take into account more than one kind of input data. For example, a multi-modal self-supervised learning (SSL) can be found in Siriwardhana's study [13]. The research extracts features from data relating to voice, face, and text. It has been claimed that the SSL approach is efficient in terms of performance in investigations that are carried out on pre-trained models. It is also mentioned that this technology can address a variety of issues, including sentiment analysis, among others. In the study, approximately 87% met with success.

3. METHODOLOGIES

In this part of the study, the adopted methodologies to create an emotion classification model are mentioned.

3.1. Neural Network Overview

In recent years, significant advancements have been achieved in the domain of machine learning through the utilization of neural networks. The concept is derived from the cognitive processes of the human brain [14]. The objective is to acquire information by investigating the learning processes employed by individuals. The system emulates the cognitive processes of neural networks seen in the human brain, specifically their capacity for learning and generalization. Artificial neurons are developed based on the communication patterns

observed in biological neurons. The aforementioned computational method is commonly employed in contemporary academic disciplines, including but not limited to language processing, audio and image categorization, energy generation, and automobile research. In addition, individuals have the capability to successfully fill in incomplete patterns. Due to its non-linear structure, this approach demonstrates efficacy in addressing intricate problems. This section presents an overview of neural networks.

A neural network consists of several layers and neurons. Additionally, these layers are comprised of nodes and artificial neurons. Neural networks typically have an initial layer for input, followed by one or more intermediate layers, and culminating in an output layer. A neural network that comprises only one layer is referred to as a single-layer neural network, while a neural network that incorporates many hidden layers is known as a multi-layer neural network. The process of information transfer takes place via the interconnections established among neurons within the various layers [15]. Layers of a neural network:

Input layer: It is responsible for receiving the data from which the features will be extracted. Each neuron within a neural network represents a distinct characteristic. Inputs refer to the information that is received by neurons and subsequently undergoes processing. The data may encompass several forms, such as images, text, or other types. That data is transmitted to the nucleus of the neuron. Typically, it is presented in a numerical manner. Predictions are generated through the execution of mathematical operations on numerical values presented in specific formats. It denotes the measure of the strength of the signal being received. The incoming signal undergoes multiplication with the respective connection weights before being transmitted to the central processing unit [16].

Hidden layers: They are responsible for carrying out intermediate computations within a neural network. The inclusion of intermediate layers inside the network architecture enhances the capacity for deeper learning.

Output layer: It is responsible for generating predictions. The quantity of neurons in the last layer is subject to variation based on the specific characteristics of the operation being

executed. While a cellular entity may possess several inputs, it is limited to a singular output point.

The network exhibits bidirectional feeding. To clarify, the process of forward and reverse feeding is characterized by the repetition of the cycle. Advancement is achieved by traversing the network and generating predictions. The back-propagation algorithm facilitates the reevaluation of gradients. By recalculating gradients, the weights are revised, thereby enhancing the outcomes. The utilization of the activation function is employed in these processes.

The activation function can be described as a mathematical procedure. The output of the neuron is determined. The summation of weights at the input of the neuron is followed by a mathematical operation. The utilization of an activation function enhances the network's capacity for learning. It also has a significant impact on the performance of the neural network. The neural network aggregates the weights from preceding layers to provide an output value. There are various activation methods that are accessible [17].

The sigmoid function is responsible for computing the output within the interval of 0 and 1.

The hyperbolic tangent function, denoted as tanh, is a mathematical function that calculates the output inside the range of -1 and 1.

The Rectified Linear Unit (ReLU) is an activation function commonly used in neural networks. It operates by preserving positive input values and setting negative input values to zero.

The linear function is one that outputs the input without undergoing any form of processing.

The figure 3.1 illustrates the procedure of the calculating weight of a layer [15]. The weights of each input in the layer are multiplied by their respective values and then summed. Next, the bias value is incorporated. Ultimately, the resultant value is passed on to the activation function. The computed value is transmitted to the subsequent neuron.

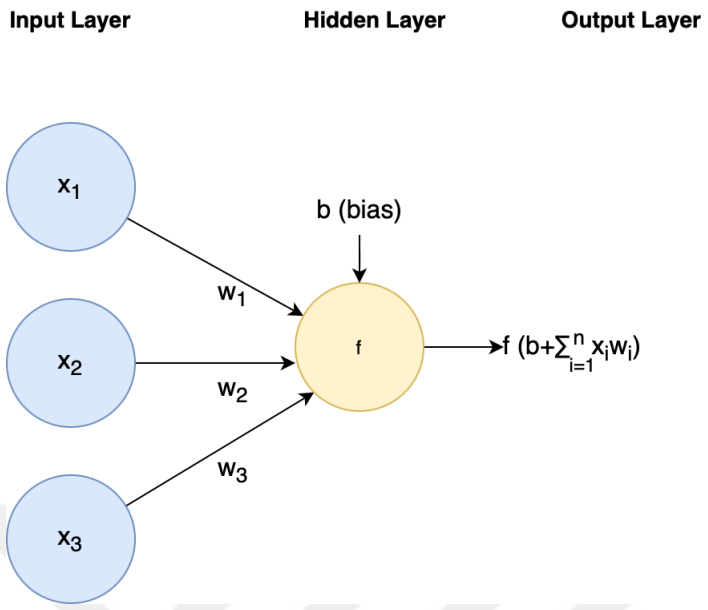

Figure 3.1 Neural network weight calculation.

Various architectures can be employed in the development of neural networks. Consequently, investigations are conducted within several subcategories. The following are examples of subgroups within the field of neural networks.

The Single-Layer Perceptron (SPL) is composed of a solitary input layer and a solitary output layer [18]. Linear functions are commonly employed in various academic disciplines. It has application in simplistic categorizations.

Multi-Layer Perceptron (MLP) consists of one or more intermediary layers. Non-linear functions may be employed [19].

The Recurrent Neural Network (RNN) is a type of neural network that is specifically designed for processing sequential data [20]. This method is employed for the analysis of temporal dependencies. The data from the previous phases is preserved. Retaining of this information has an impact on subsequent outcomes. This technique is employed in the study of temporal data.

Long Short-Term Memory Networks (LSTM) are developed to address the challenges encountered in Recurrent Neural Networks. This method involves the storage of time-based

information. The system consists of three distinct components: a forget gate, an input gate, and an exit gate [21]. The determination of which details are to be forgotten occurs at the threshold of forgetting. The information that will or won't be collected at the entrance gate is examined. The exit gate is responsible for determining the information that will be transmitted to the subsequent input. Temporal data is also used in this network.

In addition to the preceding subgroups, there exist research endeavors pertaining to neural networks within the domains of Convolutional Neural Networks, Operational Neural Network (ONN), and Spike Neural Network (SNN). These topics are covered in more detail in the following subheadings. Since this study employes these three methodologies.

3.2. Convolutional Neural Networks

Convolutional neural networks have demonstrated efficacy in tasks that heavily depend on visual data and are commonly employed in many computer vision applications, such as image classification, picture recognition, and object detection. In order to differentiate between photos, CNNs analyze several features present within the image. As an illustration, a model trained on a data set containing cats can differentiate between cats by acquiring knowledge of several distinguishing characteristics, including their four-legged nature, ear size, and facial feature clarity. CNN acquires knowledge through the identification and exploration of distinct tangible characteristics inherent in these items. Indeed, the human brain is capable of segregating items in such a manner. However, it is possible for humans to engage in this behavior without conscious awareness. The CNN architecture comprises 6 distinct layers [22].

- Input layer
- Convolutional layer
- Activation layer
- Pooling layer

- Flattening layer
- Fully-connected layer

Convolutional layers are a fundamental component of convolutional neural networks, where pictures are represented as matrices. The matrices contain the pixel values corresponding to the photos. Within this particular stratum, many characteristics are derived and isolated from the given image. Multiple filters are employed to manipulate the photos. Various image processing techniques, such as edge detection or corner detection, can be employed to implement filters. Matrices undergo the application of filters. Typically, filters are commonly found in dimensions of either 3x3 or 5x5. The application of filters is performed on the matrix representing the image. The dot product operation is performed on the image matrix, with the filter matrix being applied to it. This process begins at the upper left corner and proceeds by shifting to the right. The previous method is iteratively executed, beginning from the lowest row of the image matrix, and thereafter extended to encompass all constituent elements inside the matrix.

Upon the completion of the multiplication process, a resultant matrix is generated. The dimensions of the resulting matrix are determined using the equation (1).

$$n_{out} = n - f + 1 \tag{1}$$

where n is the image matrix size and f is the filter matrix size

Features are extracted by employing filters on the photos. CNNs can be utilized for image processing. Therefore, it is possible to enhance its appeal through the incorporation of further characteristics. In typical circumstances, the filter matrix undergoes a displacement of one pixel dimension as it traverses the picture matrix. However, it is possible to alter this value. The practice of shifting steps is commonly referred to as ”stride” [22]. When the stride value is set to 2, the filter matrix progresses by traversing the image matrix while skipping 2 pixels at a time. When the value of the stride is not equal to one, the size of the output matrix is

determined using the equation (2).

$$n_{out} = \frac{(n - f)}{s} + 1 \tag{2}$$

where n is the image matrix size, f is the filter matrix size and s is the stride number.

In order to get an output matrix that is the same size to the input matrix, additional pixels are appended to the input matrix. The aforementioned procedure is commonly referred to as padding [23]. In this procedure, the dimensions of the input matrix are expanded by appending zero-valued elements around the perimeter of the original matrix.

Activation layer is applied following the convolutional layer to introduce non-linearity to the network. In the absence of network activation, a linear model will be implemented. In order to acquire diverse information, it is necessary to incorporate an activation function within the model. The output generated by the activation function is subsequently transmitted as input to the subsequent neuron [23]. Various activation functions can be employed based on the nature of the problem being addressed.

Pooling layer is responsible for implementing the reduction procedure [24]. The dimensions of the matrix have been decreased. Therefore, it decreases the complexity of calculations and enhances overall performance. Additionally, it mitigates the potential occurrence of overfitting. There exist two distinct forms of pooling techniques: average pooling and max pooling. The process of average pooling involves the creation of a novel matrix through the computation of the average value of the pixels included by the pooling filter. The process of max pooling involves generating a new matrix by selecting the highest-valued pixels within a given pooling filter.

Flattening layer is responsible for transforming the input matrix into a suitable format for the Fully Connected layer, which is the final layer in the network. The matrix originating from the convolutional and pooling layer is converted into a one-dimensional structure [24].

Fully-connected layer is responsible for executing the categorization task. The data is received at this layer on a singular plane [23]. Artificial neural networks receive incoming

data as their input. The inputs are multiplied by a weight corresponding to their association and then combined with the bias value by summation. Subsequently, it is transmitted to an activation function. Hence commences the process of acquiring knowledge. Subsequently, the categorization outcome is acquired.

Convolutional neural networks provide the capability to extract significant characteristics and acquire diverse information from intricate visual representations. The reduction in the number of parameters is achieved by the sharing of calculated weights across filters. The occurrence of learning becomes more effective. The technology has been successfully applied in various domains, including traffic surveillance for license plate recognition, camera systems for facial recognition, and medical imaging for disease detection within the healthcare sector.

3.3. Spike Neural Network

Nowadays, various deep learning models are produced. The aim of these new models is to find a more effective calculation method. One of these new studies is spike neural networks [25]. The advent of this brain inspired method addresses the disparity in energy efficiency in existing models and enables the development of better designs.

The objective of neuromorphic engineering is to decrease the energy consumption of deep learning systems. It is based on the principles of the brain. Neuromorphic sensors take influence from biologic systems like cochlear and encode the signals as spikes. Traditional models employ continuous values for the purpose of extracting characteristics from information. In contrast, spike neural networks utilize the transmission of knowledge through spikes [26]. SNN is a neuromorphic method and uses single-bit operations. Spike-based algorithms are used in various disciplines, including robotics and medicine.

Spikes, sparsity, and static suppression are important notions within the field of neural networks. Electrical impulses, sometimes referred to as spikes, play a vital role in the processing and communication functions of biological neurons [25]. Single-bit events are frequently simplified, resulting in a decreased requirement for high-precision activation and

a reduction in carry propagation latency. Using spike-based methods only needs to multiply a weight by a spike, which makes it easier to read the weight number from memory. Instead of using binary numbers, spike networks use clock data spread across a digital device to represent time, which is not a binary number. Sparsity is a significant characteristic observed in biological neurons. Since they primarily remain in a resting state, resulting in a reduction of the majority of activations to a value of zero. Sparse tensors offer a more affordable storage solution due to their ability to occupy storage only based on the amount of non-zero components. This characteristic effectively reduces the space needed for basic data structures. Static suppression is a physiological process within the sensory system that enhances the activity of neurons when they are subjected to dynamic events while concurrently inhibiting their response to static data. This phenomenon transpires over brief temporal intervals and can also manifest over extended temporal intervals, like seconds.

The voltage leakage is described by SNNs using the Leaky-Integrate-and-Fire (LIF) model [27]. Its structure is not linear in nature. It is a single-layer structure, consisting solely of an input and output layer. The neuronal network utilizes loops or multidirectional connections to transmit data. As a consequence of this detailed architecture, diverse learning algorithms are necessary. Adjustments need to be made to the backpropagation methods. When data does not arrive simultaneously, artificial neural networks prove to be inadequate.

Time-streamed data may be processed using SNNs. The development of spikes is affected by distinct functions. Not all neurons in the model exhibit firing activity throughout each cycle. The inputs are subjected to a calculation process in order to determine a certain weight value [28]. The membrane potential is expressed as the sum of all of the variables. This value serves as a threshold. If the specified threshold value is not met, the operation will not be executed. Actions are executed at values that are equivalent to or higher than the threshold amount as shown figure 3.2 [25].

LIF threshold equation is shown in equation (3).

$$I_{inj} = C_m \times \frac{dV(t)}{dt} + \frac{V(t) - V_0}{R_m} \tag{3}$$

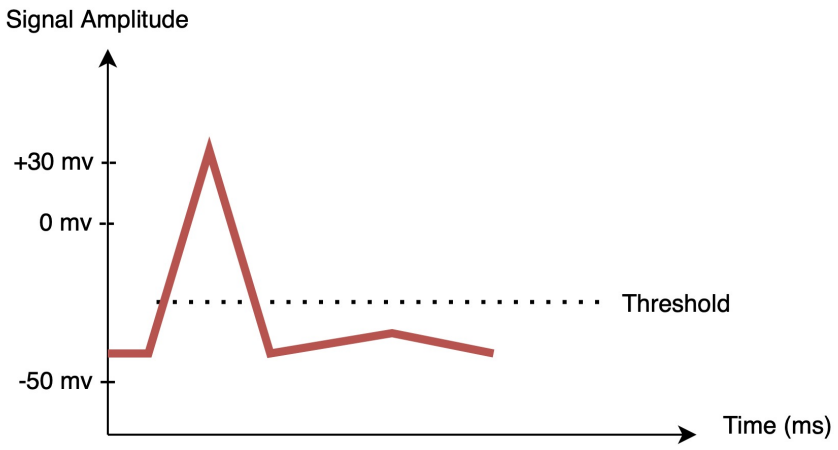

Figure 3.2 Signal amplitude distribution of a neuron.

C_m and R_m represent the membrane capacitance and resistance, correspondingly. The membrane potential of the LIF model is denoted as V, whereas V_0 represents the resting potential. I_{inj} refers to the current that is injected into the neuron.

3.4. Operational Neural Network

An operational neural network is another type of neural network model that has been constructed to improve the weight calculation technique's. This network also consists of several layers. Each layer of the system analyzes the data and uncovers novel characteristics. Its approach exhibits several disparities when compared to conventional approaches.

ONNs are designed based on the principle that biological neural systems, such as the visual systems in creatures, consist of diverse, non-linear neurons with different connections between them [29]. When compared to the traditional linear equivalents, the operational neural network is claimed to have better learning performance. The objective of our work is to show the impact of ONN techniques on classification comparing the traditional neural network techniques. We also to address the issue of emotion categorization by developing a model that is based on the ONN architecture.

A traditional convolution is computed by performing element-wise multiplication between the input and the kernel matrix and then summing the resulting values. The overall weight is

determined by multiplying the value of each input by a specific weight [30]. Afterwards, a scrolling process extracts characteristics from the data. The completed output matrix is then moved to the subsequent layer after passing through an activation function like ReLU.

The weight calculation of traditional convolution is represented mathematically by the equation (4).

$$y = f(\sum_{i=1}^{n} w_i x_i) \tag{4}$$

The distinguishing factor of the operational neural network is the exclusion of linear functions. Instead of using linear functions, nodal ψ and pool ϕ procedures are employed. Input and weight matrix operations are executed in the following manner.

The weight equation of ONN below demonstrates the input-weight operations in the operational neural network. The calculation of weights is performed using a nodal and a pool operator rather than the summation operation of the products of the weights and inputs as stated in the weight calculation of traditional convolution. Following these procedures, it is then transmitted to the activation function.

$$\sum_{i=1}^{n} \phi(\psi(w_i x_i)) \tag{5}$$

Pool and nodal operations can be chosen differently depending on the needs of the problem to be solved. The tables below present different types of nodal and pool operations [31].

Table 3.1 Nodal operators representations.

Generic	$\psi(w, x)$
Sinusoid	sin(wx)
Exponential	e^{wx}
Multiplication	wx
Chirp	sin(wx^2)

Table 3.2 Pool operators representations.

Generic	$\phi(S_l)$
Median	median(S_l)
Summation	ΣS_l
Max	max(S_l)

The weights are modified based on these procedures. In summary, the nodal operator processes each weight and input, and the results of these operations are combined to generate the final output of the receptive field.

Functional neural networks have applications in several domains. It has the capability to function in domains such as picture and sound identification as well as natural language comprehension. Patterns can be identified in intricate data sets within these domains. ONN is employed as one of the deep learning techniques in this study.

3.5. Pre-trained Neural Networks

The field of machine learning has witnessed significant advancements, leading to the discovery of many methodologies in data analysis. One of the methodologies employed is the utilization of pre-trained models. Pre-trained models refer to models that have undergone training using extensive data sets. These models are trained to recognize intricate patterns through the utilization of data sets [32]. Additionally, it provides many benefits. In contemporary times, a multitude of challenges can be effectively addressed by employing these methods.

Pre-trained models are constructed using a multi-layered architecture. The model undergoes training using a large data set consisting of millions of data points. Subsequently, the models have been provided with comprehensive capacity. Typically, training involves utilizing data sets including many sorts of data, such as images, audio, and text. The process involves instructing the model using a set of pre-existing data that has been carefully selected and

providing it with a certain pattern. The primary objective of these models is to address intricate problems by leveraging extensive data sets.

During the training process, the model acquires knowledge by adjusting its parameters, commonly referred to as weights. The system effectively accomplishes certain tasks by utilizing weight factors derived from its intricate architecture and training on large-scale data sets [33]. Weights are symbolic representations of particular patterns. Neurons are responsible for the processing of incoming data and the subsequent generation of an output. The specifics of these procedures are determined by the weights.

Pre-trained models have certain benefits. They successfully acquire the general characteristics and patterns of the issue because they have been trained on massive data sets. The utilization of predetermined weights demonstrates efficacy in terms of performance. The training process exhibits a high degree of speed. This enables the model to more rapidly adjust its parameters in order to effectively address the challenge at hand. Additionally, it provides the chance to engage with limited data sets. Given the existing weight settings, it is possible to attain favorable outcomes even when the tested data set has a restricted amount of data.

Pre-trained models are utilized across several domains. Sentiment analysis, namely in the form of text classification, has demonstrated its efficacy in addressing certain challenges. The pre-trained picture classification model of this technology finds application in the field of object recognition. For instance, security cameras can yield outcomes in the classification of diverse items. In the domain of healthcare, investigations can be conducted to ascertain discoveries using diverse medical imaging techniques. Therefore, a range of disorders can be identified and diagnosed.

The analysis of objects observed using sensors employed in autonomous cars is feasible. The vehicle has the capability to undergo training based on environmental circumstances. Within the realm of marketing, content suggestions may encompass the provision of recommendations to users as well. Financial data can be utilized for conducting market analysis. Pre-trained models have the potential to be utilized across various domains.

The following is a compilation of well-known pre-trained models:

ResNet, also known as Residual Networks, is a deep learning architecture that has gained significant attention and popularity in the field of computer vision. Deep learning employs this technique for the purpose of classifying images and objects [34]. The process involves establishing a connection between the input and output of a given block. The aforementioned connections are referred to as Residual Connections. These blocks facilitate the acceleration of network training. It provides more comprehensive and inclusive generalizations.

The Generative Pre-trained Transformer (GPT) is a state-of-the-art language model that has been pre-trained on a large corpus of text data [35]. This technology has applications within the domain of natural language processing. The model undergoes training using extensive text data sets and acquires knowledge of diverse patterns in language structure. Text classification, language translation, and sentiment analysis are among the various applications where it finds utility. The application of this technology extends to various domains, including but not limited to language comprehension, language manipulation, categorization of textual data, and generation of text segments.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a widely employed model in the field of natural language processing. This model possesses the capability to comprehend and analyze texts from multiple perspectives [33]. The process involves examining the preceding and subsequent words with which the words are associated in order to derive semantic significance from the sentence. It offers assistance in language translation and responding to inquiries.

MobileNet model is specifically designed and optimized for deployment on mobile devices [36]. The purpose of this technology is to facilitate the identification and classification of objects. This technology is employed in gadgets that have restricted performance capabilities. Mobile devices are the primary platforms on which it is utilized. The low memory utilization of the application results in a high level of performance on mobile devices. The architecture provided exhibits computational efficiency. It employs a reduced number of parameters.

Another often used pre-trained model is the Visual Geometry Group (VGG16). The VGG16 architectural approach is employed to generate a model in this study.

Visual Geometry Group is a research group that focuses on the study of visual geometry. This technology is employed for the purposes of object recognition and classification. The development of this technology was undertaken by academics affiliated with Oxford University. The purpose of this system is to derive meaning and categorize information by extracting semantic representations from intricate visual stimuli. The architecture is composed of a total of 16 layers. The algorithm performs a classification task involving a total of 1,000 distinct classes. The model is comprised of 138 million parameters. The architectural framework of the construction can be described as follows [37]:

- 13 convolutional layers
- 3 fully connected layers
- ReLU (Rectified Linear Unit) for activation function
- Max pooling layers

VGG-16 architecture is shown in figure 3.3.

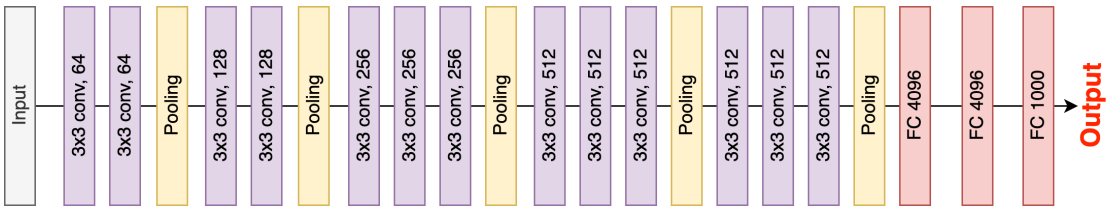

Figure 3.3 VGG16 structure.

4. MULTI-INPUT OPERATIONAL NEURAL NETWORK APPROACH

4.1. Overall Methodology

The objective of the study is to create an advanced deep learning model for the purpose of accurately classifying emotions. Numerous studies on this subject can be found in the existing literature. Typically, studies involve the creation of models using data sets that consist of either images or text. The aim of this study is to create a model using audio-visual data sets to classify emotions. The rationale behind selecting data sets that include video is to enable comparisons with models that extract features solely from image or audio data. As previously stated, there are certain issues with the classification of emotions. The models in the literature may have certain deficiencies as a result of the uncertainty behind the experienced emotions. Features obtained solely through text or images may lack certain functionalities. Therefore, this study develops a model with two inputs. The objective is to acquire a greater variety of characteristics in the two-input model.

The model is constructed using ONN. To observe variations in the effectiveness of ONN across the other neural network types, additional models are constructed using SNN and CNN. Therefore, an efficacy comparison of ONN with other neural networks can be made.

The model has undergone training using the video data set. Before being provided to the model as input, videos go through a number of processing steps. Following these procedures, the model generates two inputs. Both of these entries are in the form of images. The first input involves the selection of a frame from the video. Visual features are extracted from the videos by selecting the video frame. Prior to selection, the frames in the videos are carried out analysis to verify that the chosen frame actually represents the crucial frame in the video. The videos undergo the processes to get the main frame as the first input described in the following section.

4.2. Key Frame Selection

A video is comprised of a series of frames. The interpretation of a video can vary depending on each individual frame. Given that each frame has the potential to possess distinct visual characteristics. Different impacts can be observed in the model if different video frames are transferred to it. The selection of a frame must prioritize the identification of superior solutions in comparison to other frames. Therefore, the process of key frame selection is implemented in the videos. The process of key frame selection is employed in a range of multimedia and video-oriented applications with the aim of extracting a limited group of frames from a video sequence. These frames are often chosen to represent events that are deemed noteworthy or crucial. The essential frames, which have been carefully chosen, serve as a concise depiction of the video's content, facilitating the process of previewing, analyzing, and navigating through the video. In order to guarantee the precision and efficiency of the extracted frame, it is important to ascertain its accuracy and compactness. The process of key frame selection can be summarized as follows:

- Analysis of the differences in LUV values across the frames
- Evaluation of the brightness score difference among the frames
- Entropy filtration
- K-Means clustering
- Comparison of Laplacian variances between frames

To begin with, frames are extracted from the video. The frames that have been obtained are depicted using the LUV color space. The LUV color space is employed in the fields of color science and computer graphics to accurately represent colors in a manner that is both perceptually uniform and independent of the specific device being utilized. The LUV color space is comprised of three components. The variables denoted as L (Luminance), u, and v (Chromacity) are referred [38]. The luminance value, denoted as L, is used to

quantify the brightness of a color. The variables u and v are utilized for the purpose of storing color data. The u component denotes the spatial location along the red-green axis, while the v component signifies the spatial location along the yellow-blue axis. The LUV color space is chosen because of its characteristic of perceived homogeneity. To clarify, a distinct dissimilarity in LUV values denotes a reliable and anticipated dissimilarity in seen color throughout the entirety of the color space, as opposed to alternative color spaces such as Red, Green, Blue (RGB) and Cyan, Magenta, Yellow and Black (CMYK).

During the second step of key frame selection, the absolute LUV color differences of each frame are determined. The discrepancies between the LUV values of the current frame and the preceding frame are compared. The frames exhibiting significant disparities are collected. Multiple distinct frames are extracted from this process. The difference formula is based on Euclidean distance [39]. In the context of LUV color space, the distance between two data points (L_1, U_1, V_1) and (L_2, U_2, V_2) can be determined using the calculation in the equation (6).

$$\sqrt{(L_2 - L_1)^2 + (U_2 - U_1)^2 + (V_2 - V_1)^2} \tag{6}$$

When the distance between two points decreases, the similarity of their colors increases. Conversely, an increase in disparity signifies a divergence in color.

Subsequently, an analysis is conducted on the brightness ratings of each filtered frame. The luminance value of a frame is indicated by its brightness score [40]. The relationship between brightness and pixel intensity is one of direct proportionality. The intensity of each pixel is computed. The brightness score of each frame is likewise subjected to comparison. Next, the entropy of each filtered frame is calculated. Entropy is a measure that quantifies the level of uncertainty or randomness [41]. It is associated with a pixel inside a frame. Following the analysis of entropy, a subsequent step involves the implementation of contrast filtering on the frames. The term ”high contrast” refers to an image that has distinct and well-defined edges. The process of clustering filtered frames involves the utilization of the K-Means algorithm. The K-means algorithm partitions the elements within a given data set along the coordinate axis based on the specified number of K clusters [42]. Elements that possess a

certain degree of similarity are grouped together in close proximity. Hence, the identification of commonalities among locations is observed. Histograms of each frame are utilized in the clustering procedure.

The selection of the optimal frame from the clusters is achieved by employing the variance of the Laplacian sorting technique subsequent to the clustering process. The Laplacian method variation is frequently employed in the analysis of photographs to detect instances of blurring. The methodology demonstrates the distribution of pixel values. The algorithm assesses the number of features and frequencies present in the image as well as identifies locations where pixel values exhibit rapid changes, such as along edges or within small-scale features.

The value is determined by the summation of the second derivatives of the pixels within a two-dimensional Laplacian image. The formula is depicted in the following manner (7) [43].

$$\nabla I(x, y) = \frac{d^2 I}{dx^2} + \frac{d^2 I}{dy^2} \tag{7}$$

Let $I(x, y)$ represent a pixel in the image, and $\nabla I(x, y)$ denote the Laplacian of the pixel.

The calculation of the Laplacian value is performed across the entirety of the image. A high variance indicates that the image contains a significant amount of small detail. The sorting process ensures the selection of the least blurry image inside the cluster. The determination of image blur is based on the Laplace variance. Hence, it has the capability to ascertain whether the image is blurred or not. Ultimately, the filtered frames yield the key frame, which is subsequently employed as the first input in the model. Figure 4.1 illustrates the method of selecting crucial frames.

The key frame is chosen based on the steps listed above. The first input of the model is acquired at this point. The model's second input consists of the sound spectrogram derived from the videos. Sound spectrograms are visual depictions that illustrate the characteristics of the audio in a video. The second objective is to acquire accurate spectrograms in order to extract sound characteristics from the video and use them as second input for the proposed

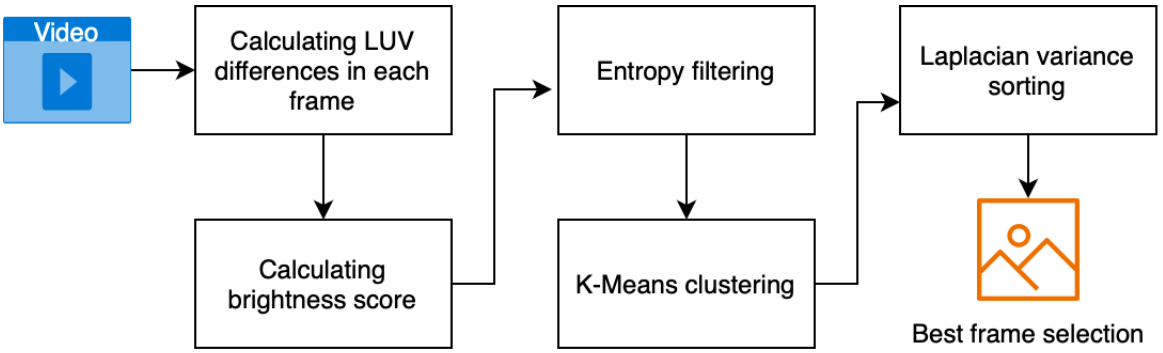

Figure 4.1 Key frame selection.

model. Multiple sound spectrograms exist. The Mel spectrogram is chosen for this study. The Mel spectrogram displays the time-dependent variation of sound signal frequencies. Comprehensive information regarding Mel spectrogram extraction from the videos can be found in the following section.

4.3. Mel Spectrogram Extraction

A spectrogram is a graphical representation that illustrates the frequency content of a signal. The provided visual representation depicts the temporal variation in the frequency of the signal. The temporal variation of frequency density is depicted [44]. Typically, spectrums are visually represented on a two-dimensional coordinate system. The temporal dimension is represented on one axis, while the frequency dimension is represented on the other axis.

There are multiple formats available for visualizing audio spectrograms. The Mel spectrogram is considered to be one of the approaches used in this context. Linear scales are commonly employed in conventional spectrograms. The Mel spectrogram represents frequencies on the Mel scale.

The Mel scale is a perceptual scale that represents sound frequencies based on the human auditory system. Individuals do not exhibit sensitivity to frequencies in a linear configuration. Sound changes are perceived more effectively at lower frequencies. The likelihood of achieving success diminishes as the frequency of attempts increases. The Mel scale has been devised with consideration for the auditory perception of humans. The scale

in question is one that effectively represents sound frequencies in a manner that closely aligns with the auditory perception of humans. The auditory system of humans does not exhibit a linear structure with regards to its sensitivity to different frequencies. The ability to notice changes is more pronounced at lower frequencies, although it diminishes as the frequency increases. The Mel scale is a scale system that has been specifically constructed to accurately represent the human sense of frequency. Spectrograms find extensive application in various disciplines, including music, linguistics, radar, and sound processing models. The presented research examines and visualizes the frequency of the auditory stimulus in these investigations. The Mel scale is employed in our research to examine the emotional content of sound frequencies within the data sets.

The spectrogram is generated through the utilization of Mel-frequency cepstral coefficients (MFCCs) or Mel-frequency filter banks on the audio input, leading to the production of a visual representation that effectively conveys significant spectrum details [45]. In the domain of audio signal processing, this particular format is commonly employed for several applications, including speech recognition and music analysis. The construction of a Mel spectrogram image often involves several common procedures, including windowing, Fourier transform, frame segmentation, compression, and picture representation. Mel spectrogram images are valuable tools for officially presenting and analyzing the spectral attributes of audio sources, effectively fulfilling this objective. Audio data analysis and machine learning operations frequently employ these tools, as their main objective is to streamline the pattern recognition and feature identification processes. Furthermore, Mel spectrogram images are commonly utilized alongside normal spectrogram images as a crucial component in audio processing and analysis. The spectrogram image obtained from the vocalization of a human being is depicted in the figure 4.2.

During this phase, the audio is initially extracted from the video. Subsequently, the Mel spectrogram graph of the sound is generated and presented as an image. The spectrogram encompasses all audio signals from the start to the end of the video. Therefore, the model's second input is generated. The key frame of the video is transferred as the first input to the model. The Mel spectrogram of the same video is transferred as the second input to the

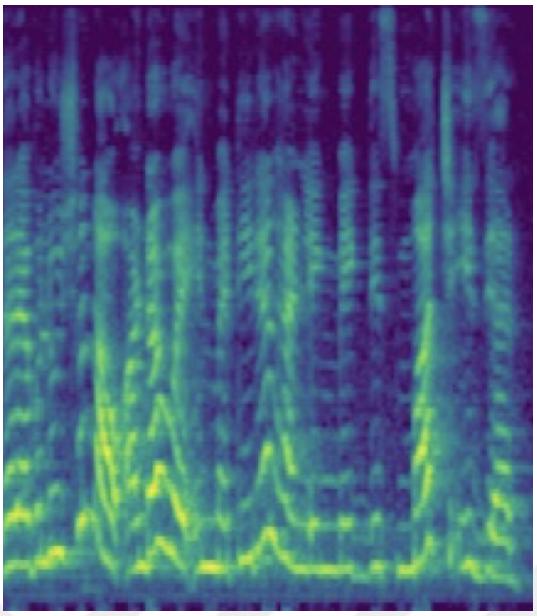

Figure 4.2 Mel spectrogram of an audio.

model. The two inputs are processed in the proposed model's layers to retrieve features. The proposed model is based on an ONN approach. In the following section, the ONN structure is explained.

4.4. Operational Neural Network Operations

Multiple two-input models are created within the scope of this study. Advancements are achieved by employing diverse neural networks for the layers incorporated in the model. The ONN methods are utilized in the initial prototype.

ONN is engineered based on the transfer mechanisms observed in real neurons. It relies on the fusion of the soma of the neuron and synaptic junctions. The nodal in the formula represents the processes of synaptic connection and pool integration [29].

Learning occurs through synaptic connections and involves a non-linear transmission process. The outputs of the neuron in the preceding layer are received as input by the subsequent layer in consecutive layers. The ONNs utilize nonlinear nodal operators such as exponentiation or sine function [31], followed by pooling the results using operators like summation or median. Activation functions are also utilized in conventional CNN topologies.

Linear operations-based neural networks may be regarded as a subset of ONN. ONN inputs consist of neural networks that incorporate both linear and non-linear connections. This guarantees a more efficient neural network. To assess the disparity in performance between the conventional neural network and ONN, we construct and evaluate our proposed model using ONN convolutional layers. Subsequently, we convert the ONN convolutional layers into standard convolutional layers and conduct further testing. In the following stages of our study, we conduct a comparison between the conventional design and the ONN architecture.

Deep learning heavily relies on the study of operational neural networks. It can also be developed. Research suggests that the general formula of an operational neural network can be optimized for GPU usage by establishing an alternative formula [31]. Furthermore, this research introduces a more expedient computational procedure. As previously stated, the ONN project showcases a novel neuron model. The process involves performing nonlinear operations on the nodal ψ and pool ϕ functions [46].

The weight equation of ONN is a generic formula. The formula can be interchanged. Initially, the input image matrix is rearranged randomly. The values are transformed into vectors and then combined to form a new matrix Y. Equation (8) displays the layer.

$$x = vec^{-}1(\sum_{i=1}^{n} Y \bigotimes W) \tag{8}$$

The symbol $\bigotimes$ represents the Hadamard product. The given calculation can be expressed as equation (9).

$$y = f_l^k(vec^{-1}(\phi_l^k(\psi_l^k(Y_{i-1}, W_l^k)))) \tag{9}$$

ϕ_l^k and ψ_l^k represent pool operations. The activation function used is the hyperbolic tangent (Tanh) function, denoted as f_l^k . The nodal function for a normal convolution operation is chosen as multiplication, represented by $\psi(\alpha, \beta) = \alpha \times \beta$. The pool function is chosen as summation, represented by $\psi(\cdot) = \Sigma$. Our model incorporates a composite nodal operator, as presented by Malik et al. [47], in which the kernels are trained to handle various powers of inputs up to a certain parameter q. This methodology relies on the extension of the Taylor

series. The Taylor series expansion for a function of order Q may be expressed in equation (10).

$$f(x)^{(Q,a)} = \sum_{n=0}^{Q} \frac{f^n(a)}{n!}(x-a)^n \tag{10}$$

The proposed formulation will provide a composite nodal operator, with the learnable parameters of the network serving as the coefficients for the powers of x. Essentially, we are doing a multiplication between various powers of inputs ranging from 1 to q, using kernel coefficients, and subsequently summing the results. The nodal operator of the k-th generating neuron in the first layer can be expressed in the following generic formula (11).

$$\psi_l^k(Y_{l-1}, W_l^k, Q, a) = \sum_{q=1}^{Q} Y_{l-1}^q \otimes W_l^{k(q)} \tag{11}$$

In the equation, the variable W_l^k represents a weight matrix with three dimensions, whereas $W_l^k(q)$ refers to a specific section or slice of the W_l^k matrix. The layers utilize the Tanh activation function f. Throughout the training phase, the weights are modified using the subsequent derivatives:

$$\frac{d\psi_l^k}{dY_{l-1}} = \sum_{q=1}^{Q} qY_{l-1}^{q-1} \otimes W_l^{k(q)} \tag{12}$$

$$\frac{d\psi_l^k}{dW_l^{k(q)}} = Y_{l-1}^q \tag{13}$$

The objective is to target a model with multiple layers. The quantity and arrangement of these layers are derived from the VGG16 architecture. The two inputs are concurrently passed to the model's layers for simultaneous processing. Essentially, the model consists of two distinct pathways that are functioning simultaneously. Trails consist of multiple layers within them. The sequence and structure of the layers in both of these pathways are identical. As previously stated, the structure of VGG determines the shape of the layers. The figure 4.3 illustrates the architectural design of the initial prototype.

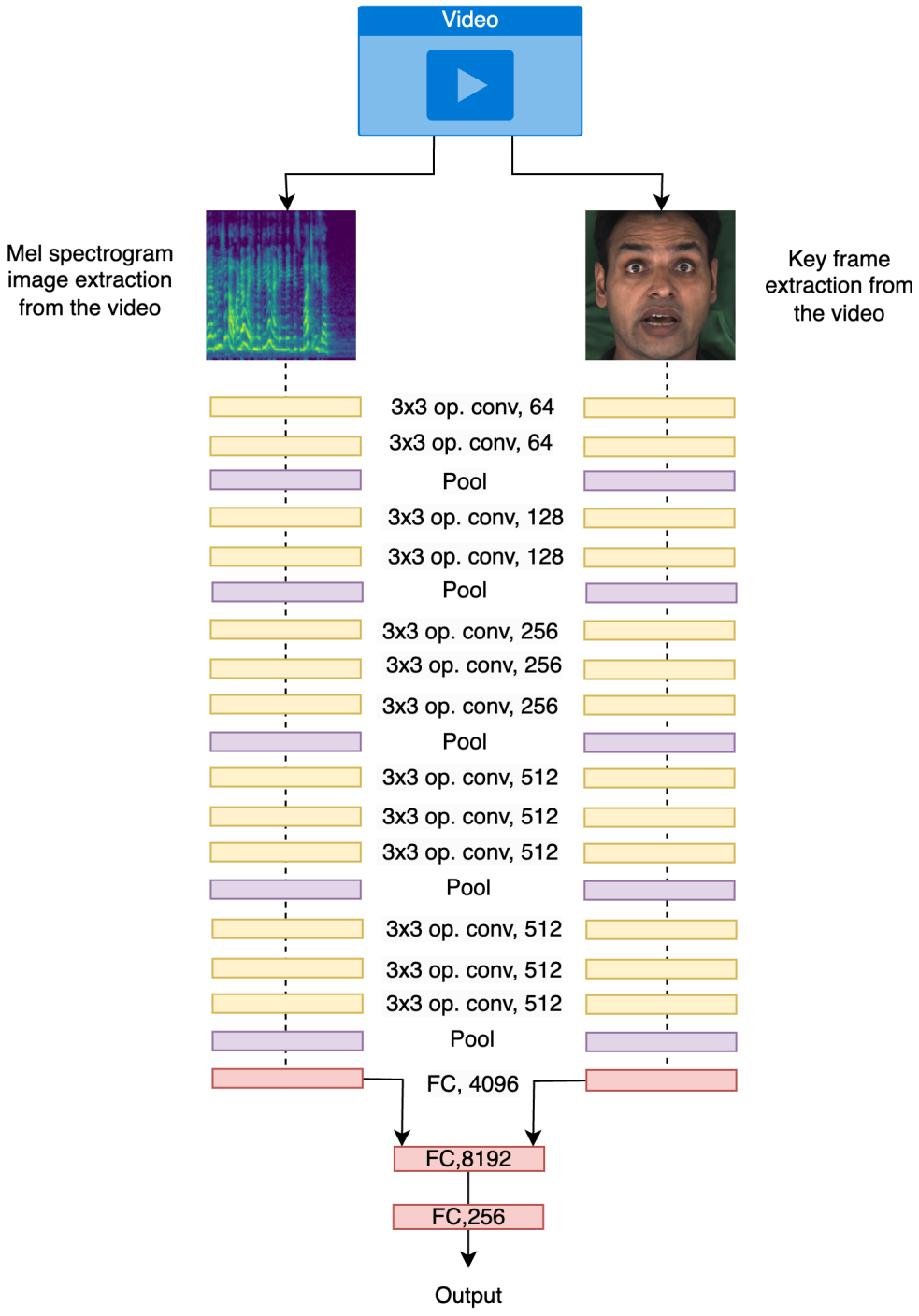

Figure 4.3 The proposed model structure.

Within a specific branch of the model, the frame extracted from the video undergoes processing. It possesses unique attributes. In the alternate branch, the Mel spectrogram derived from the audio in the video is processed to extract features. The model consists of

13 convolutional layers, similar to the architecture of VGG16. In contrast to VGG16, the convolutional layers in CNN are substituted with ONN convolutional layers. Two inputs are concurrently processed.

The model utilizes the hyperbolic tangent function (tanh) as its activation function. Once the two inputs have undergone feature extraction, the model combines the two layers. A vector with a length of 8912 is obtained. The classification process follows a linear path. Subsequently, the features derived from the frame and Mel spectrogram are combined and processed. Subsequently, the model generates a class prediction and retrieves the corresponding output. The number of classes is contingent upon the quantity of emotion classes present in the utilized data set.

4.5. Compared Neural Network Models

As stated, the proposed model mainly focuses on ONN based approaches. In order to evaluate the effectiveness of ONN on emotion classification, two distinct neural network models are created. One of them is based on convolutional neural network approaches. The other model focuses on a spike neural network. It is aimed to compare these two additional neural network models with the proposed ONN model.

4.5.1. Convolutional Network Based Model

The ONN-based model is completed training and testing using diverse data sets. In addition to the ONN model, the same data sets are tested with the CNN based model. The purpose of developing the CNN-based model is to discern the disparity between ONN and CNN. The architecture of the CNN-based model is identical to the proposed ONN-based model. The only difference lies in the convolution layers. The model that is constructed using ONN convolutional layers is substituted with CNN convolutional layers. Put simply, the original architectural framework of VGG16 is utilized. Once more, the CNN based model also consists of two inputs. The Rectified Linear Unit is employed as the activation function. The

second model, constructed using CNN convolutional layers, undergoes training and testing. It is subsequently compared to the proposed ONN based model. The figure 4.4 illustrates the architectural design of the CNN based model.

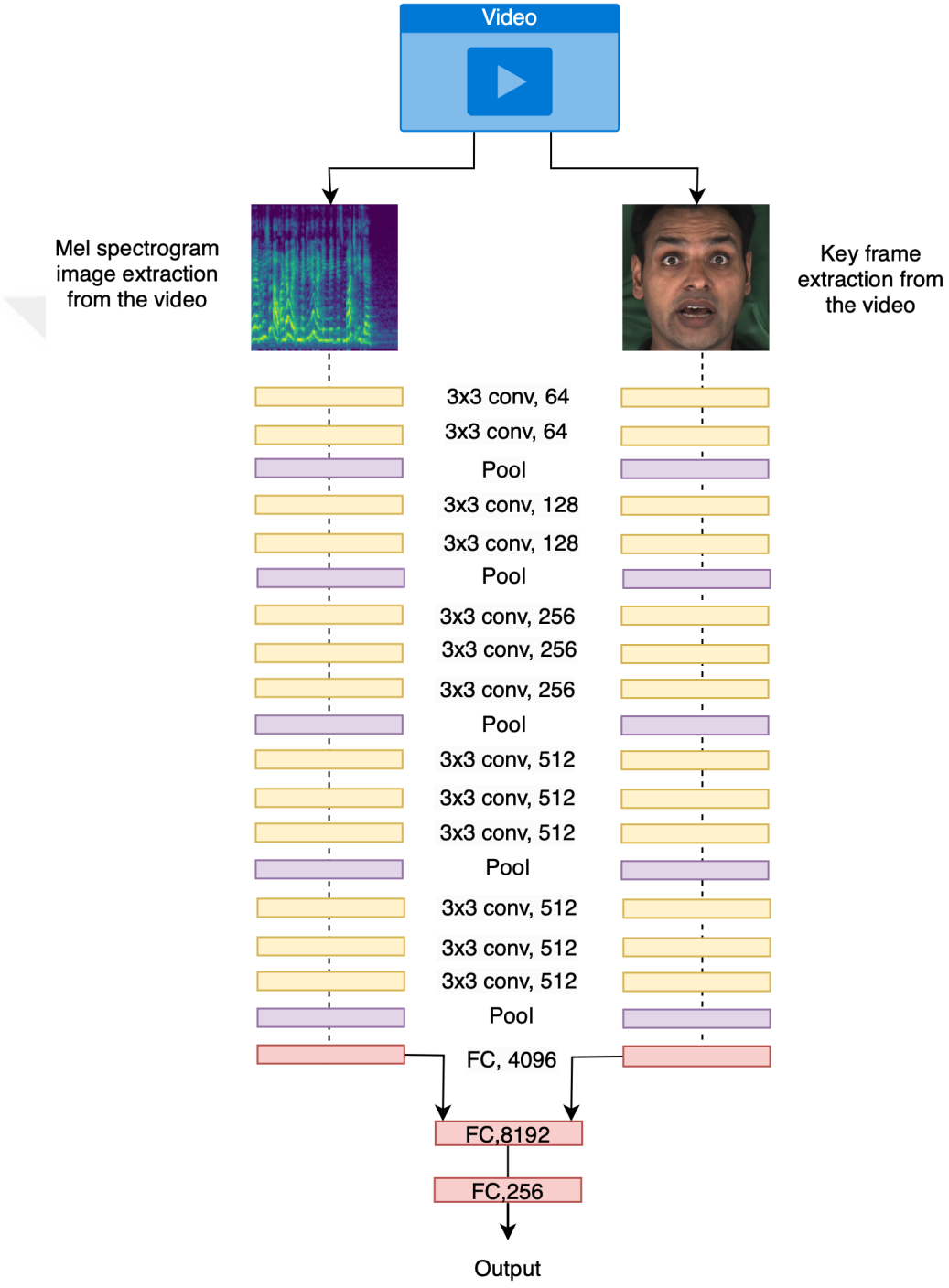

Figure 4.4 The CNN based model structure.

4.5.2. Spike Neural Network Based Model

Lastly, an additional endeavor is made to create a distinct third model utilizing SNN. As part of the SNN architecture, an effort is made to create another model that is completely separate from the CNN and ONN based models. We are conducting a study to examine the impact of various contemporary deep learning techniques on the classification of emotions. SNN is also an emerging field of study that is currently gaining significance. Furthermore, attempts have been made to employ this approach.

In SNN based approach, neurons that exceed the threshold value are propagated further, while neurons that do not exceed the threshold value are suppressed [48]. This is managed by the Spike-Time Dependent Plasticity (STDP) algorithm. STDP is a neurobiological mechanism employed by the brain to modulate synaptic connections. The strengthening of synapses is observed when triggered. The strength of untriggered synapses is reduced. The process of enhancing strength involves a boost of weight. It facilitates the temporal transmission of intercellular communication. This approach facilitates the model's acquisition of certain patterns.

A synapse serves as the connection between a pair of neurons [25]. An enhanced synaptic connection will result in increased sensitivity of the post-synaptic neuron to the resultant signals transmitted by the pre-synaptic neuron. Input spikes have the potential to produce either stimulation or inhibition responses leading to firing. When a neuron reaches a threshold level of stimulation, it will generate an action potential that propagates up its axon, transmitting signals to subsequent neurons in the network. The figure 4.5 shows the connection of two consecutive neurons. The figure 4.6 illustrates the architectural design of the SNN based model.

Synaptic weight is modulated by STDP, which is influenced by the activity of both pre-synaptic and post-synaptic neurons. Occasionally, spikes originating from both pre-neurons and post-neurons exhibit a reverse directionality upon reaching the synapse.

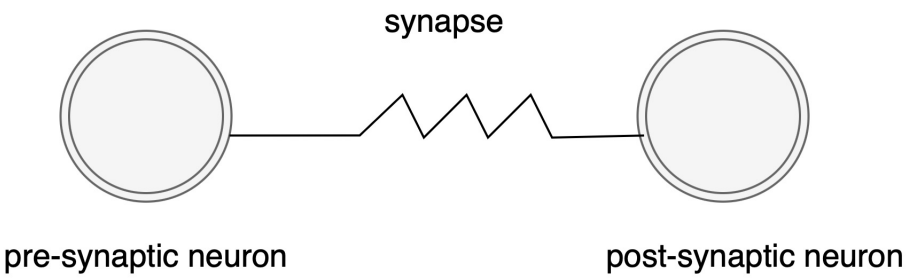

Figure 4.5 Consecutive neuron connection.

In the context of STDP, the alteration in synaptic weight is determined by the relative timing of spikes across neurons.

The equation (14) demonstrates how the shift in synaptic weight is determined by the relative timing of neuron spikes [26].

$$\triangle t(\triangle t = t_{pre} - t_{post}) \tag{14}$$

The variable $\triangle t$ represents the time difference between the arrival of a presynaptic neuron spike (t_{pre}) and the time after the spike (t_{post}). The term "t_{post}" refers to the specific moment at which the postsynaptic neuron spike occurs. When a spike from a postsynaptic neuron occurs subsequent to a spike from the presynaptic one, there is a boost in the synapse weight. When the spike of the postsynaptic neuron occurs prior to the spike of the presynaptic neuron, there is a drop in the synaptic weight.

The timing-based learning principles of STDP might provide additional complexity to the training process beyond that of typical backpropagation algorithms. Training spike neural networks might provide greater challenges compared to conventional artificial neural networks. The presence of time-sensitive neuron interactions can introduce complexities to the training process. Moreover, the utilization of spiking neural networks may necessitate the acquisition of additional hardware components. Therefore, some limitations may be experienced.

SNN based model with only one input is developed considering the operation explained above. The model is represented in figure 4.6. Upon achieving success with a solitary input,

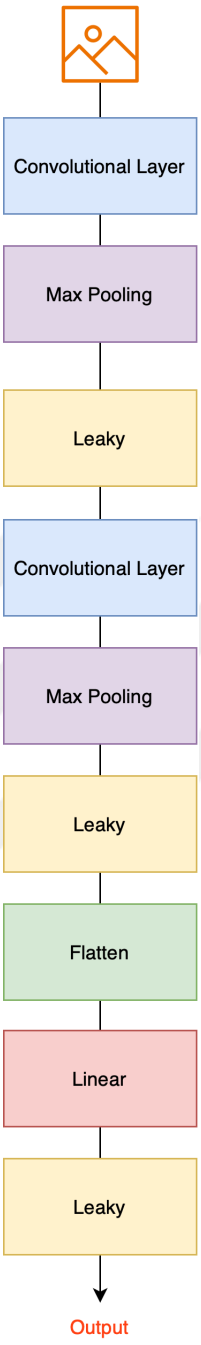

Figure 4.6 The SNN based model structure.

the objective is to construct a model with dual inputs. The input can be either a frame or a Mel spectrogram image extracted from the video. An endeavor is undertaken to construct a model using SNN. The SNN model relies on convolutional spiking neurons. The model has a 5 x 5 convolutional kernel with 12 filters. A 2×2 max-pooling function is utilized. Furthermore,

there is a fully-connected layer that establishes a mapping between 1,024 neurons and the number of emotion classes specified in the data set.

The spiking convolutional layer is converted to the Leaky activation layer. Neuronal signals undergo integration. In the absence of an input signal to the layer, a leakage develops in the neuron. A neuron's membrane potential generates an action potential when it is above a specific threshold value. Multiple training exercises are conducted. However, a comprehensive analysis has not been conducted yet. The model built on SNN needs to be further developed. This study specifically examines the ONN effect. The impact of SNN on future research can also be observed.

5. DATASET DESCRIPTION

Following the development of models for the categorization of emotions, several data sets that are trainable on these models are examined. Audio-visual data sets are required since the model analyzes both audio and visuals. When doing research on a data set, particular emphasis is placed on the picture quality of the data sets.

The analyzed data sets contain hints that include emotional expressions. Every video showcases a distinct individual. This person recites a written speech using a script. Text is meant to evoke a feeling. The individual articulates the language in accordance with the specific emotion being experienced. Consequently, a multitude of facial expressions are generated. Furthermore, the tone of voice is subject to variation based on the nature of the expression. Every video only reflects a single feeling. Videos are captured from the front-facing cameras of individuals. The people shown in the videos comprise a diverse combination of both males and females.

Within the context of this study, the process of training is conducted on three data sets:

- MEAD [49]
- Ryerson Data set [50]

- MELD [51]

These three data sets contain varying quantities of videos. The quantity of emotion categories exhibited among data sets may also differ.

The MEAD data set consists of 8 distinct emotion categories. The emotions present in this list are angry, happy, sad, contemptuous, neutral, surprised, fear and disgusted. This collection typically includes video recordings from seven distinct perspectives. However, for the sake of this investigation, we exclusively utilize photos captured from the frontal perspective of the subject. The presented emotion in each of the data is developed with performance feedback and correction. So that they guarantee the authenticity of the performed emotions. The participants are guided to talk by a team headed by a professional actor. The consistency of emotions in each data is determined to ensure the quality. A total of 1324 movies from this data set have been used in the study. The visual below displays a selection of photographs from the MEAD data set. The data set provides illustrations of 8 distinct emotion categories, as seen in the figure 5.1. The Ryerson Data set is the second data set utilized. This data

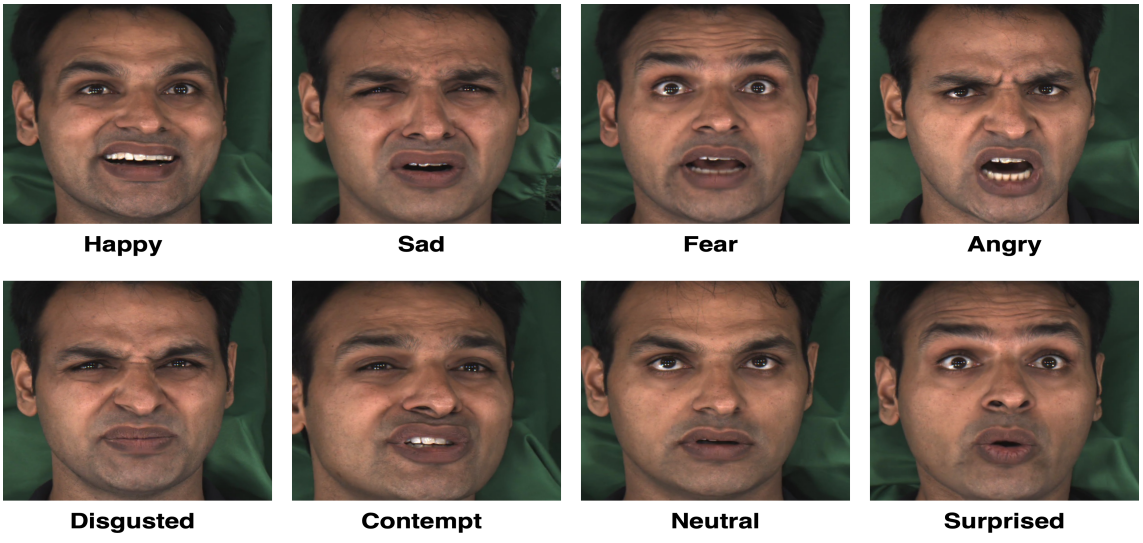

Figure 5.1 MEAD visualization.

set contains a total of 8 distinct emotion categories. The emotions include calm, angry, fear, disgusted, neural, happy, surprised and sad. Actors have the ability to employ any techniques they have been taught in to convey emotions during their performances. The

actors performed their performances without any temporal limitations. The researchers evaluate the statements. The study utilizes 540 data points from this data set. The figure 5.2 displays a selection of photographs from the Ryerson Data set. The most recent data set

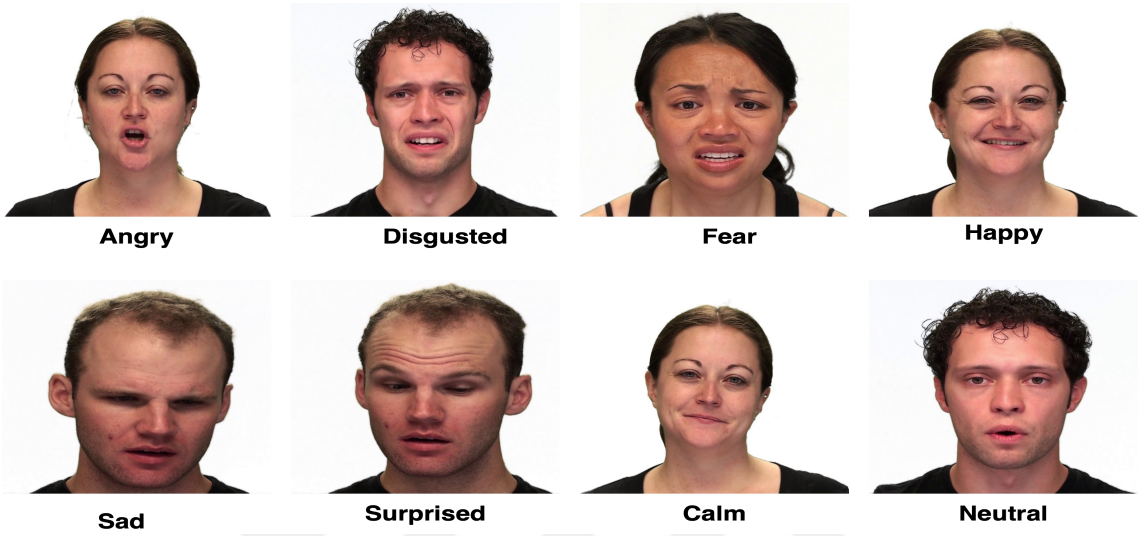

Figure 5.2 Ryerson visualization.

utilized is MELD. This data set contains instances of 7 distinct categories of emotions. The emotions are disgust, anger, fear, sad, joy, surprise and neutral. The statements in the data are labeled with the most suitable emotion group. Ekman's universally recognized emotions are utilized as annotation labels. Every statement is annotated by professional researchers. The emotions are determined by voting for each statement. The data set has a total of 470 samples. The figure 5.3 displays several samples from the MELD data set.

6. EXPERIMENTAL RESULTS

This part presents the details and outcomes of the experiments conducted on the models. The objective of the study is to categorize emotions based on audio-visual data, as previously stated. The initial development of our emotion categorization model is based on ONN approaches. The model's structure is constructed using the VGG16 architecture. This model consists of two inputs. Both of these entries are in the form of images. The model is trained

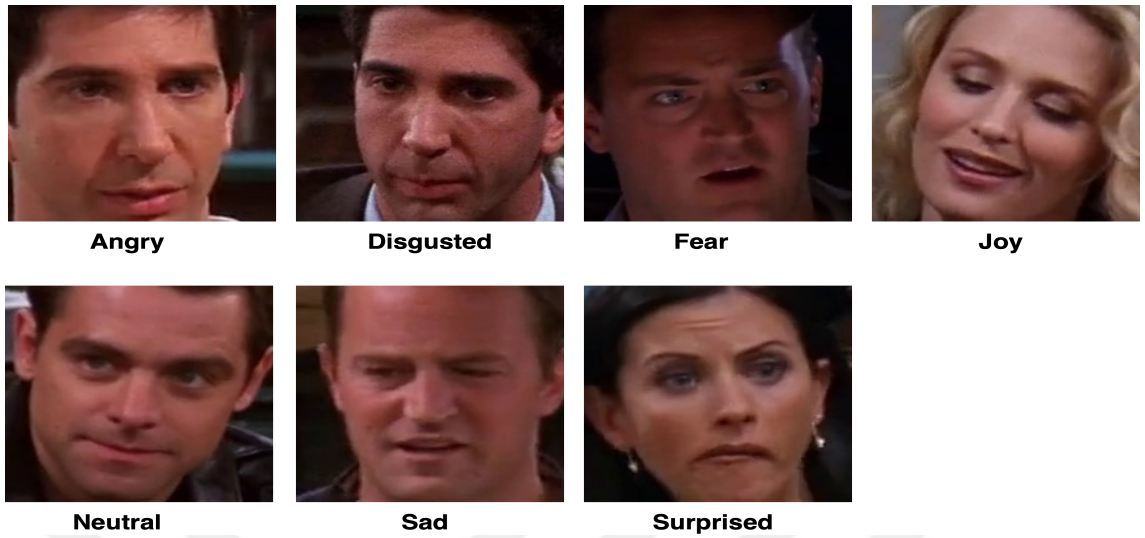

Figure 5.3 MELD visualization.

using three distinct data sets. Every data collection comprises audio-visual data. The data sets allocate 70% of the data for training, 20% for testing, and 10% for validation.

During the training process, the data from the training data set is sequentially fed into the model. A key frame is chosen from the transmitted video. Furthermore, a picture representing the Mel spectrogram graph of all the sounds in the video is generated. The model receives these two photos as distinct inputs. Features are derived from both of these input layers, which operate simultaneously. Subsequently, the characteristics are combined to generate predictions inside the classification layers of the model. Each of these three data sets undergoes separate training. Following the training process, the results are acquired by executing operations on the model inside the designated areas for testing and validation. The number of epochs in transactions is set at 80.

We employ certain metrics in our study to quantitatively assess the model's performance throughout the trial. Performance indicators play a crucial role in the analysis of deep learning models. Performance measures are utilized to quantify the efficacy and success of a model. Therefore, the determination of the model's consistency may be ascertained. Diverse performance indicators are employed in the study to enhance the evaluation of the outcomes. Evaluations encompass metrics such as classification accuracy, F1-score, precision, and recall. Precision is the quantification of the proportion of positive predictions

that are accurately positive. Recall also known as sensitivity, measures the proportion of true positive cases among all positive events. The F1-score is calculated as the harmonic mean of accuracy and recall.

6.1. Multi-Input ONN Model Results

In this phase, we provide the outcomes of training and testing three data sets using the ONN model. These findings are both stated and visually represented. The study's success is measured by displaying the findings of the aforementioned performance measures below. In addition to performance measures, we assess the effectiveness of the proposed model by comparing its results with those of previous research in the literature. The models chosen for comparison from the literature utilize the same three data sets that we employed. By using this approach, a more reliable and accurate comparison may be conducted.

The ONN model is initially trained using the MEAD and subsequently evaluated. Next, the test data set is examined using the trained model. The prediction outcomes of the MEAD are displayed in table 6.1. Based on the findings, it can be concluded that confusion may arise in those enduring anger and fear.

Table 6.1 The results of the model with MEAD.

Emotion Class	**Precision**	**Recall**	**F1-Score**
Angry	1.00	0.87	0.93
Contempt	0.93	1.00	0.96
Disgusted	0.92	0.96	0.94
Fear	1.00	0.86	0.92
Happy	0.94	0.94	0.94
Neutral	1.00	1.00	1.00
Sad	0.94	1.00	0.97
Surprised	0.92	1.00	0.96

There are three validated studies on MEAD available in the literature that may be used to compare our findings with other research [52]. These three studies are also trained using the MEAD. The initial analysis exclusively utilizes the audio data from the data set. The accuracy is reported as 40.4% with a F1-score of 0.39. The second research just utilizes pictures from the MEAD and reports its accuracy as 50.6% with a F1-score of 0.51. The third research further assesses the audio and demonstrates a 76% accuracy rate with a 0.75 f1-score. Our work utilizes both auditory and visual data from MEAD for training purposes, as previously stated. The total accuracy of the system is 96%, with a F1-score of 0.95. The findings demonstrate that our emotion categorization model effectively collected audio-visual input, as indicated in table 6.2.

Table 6.2 The comparison of the proposed model with studies in the literature using MEAD.

Model	**Accuracy (%)**	**F1-Score**
Audio based 1 [52]	40.40	0.39
Visual based [52]	50.60	0.51
Audio based 2[52]	76.00	0.75
Proposed model	96.00	0.95

The confusion matrix of the model tested with MEAD is shown in figure 6.1.

The ONN model employed is trained using the Ryerson Emotion Database. Next, the test dataset is examined using the trained model. The forecasted outcomes of the Ryerson Emotion Database are displayed in table 6.3.

Furthermore, several researchers have constructed models utilizing the Ryerson Emotion Database. The outcomes of these models are displayed in the table 6.4. The result of the proposed model using the Ryerson Emotion Data set is also presented in the table.

The confusion matrix of the model tested with the Ryerson Emotion Database is shown in figure 6.2.

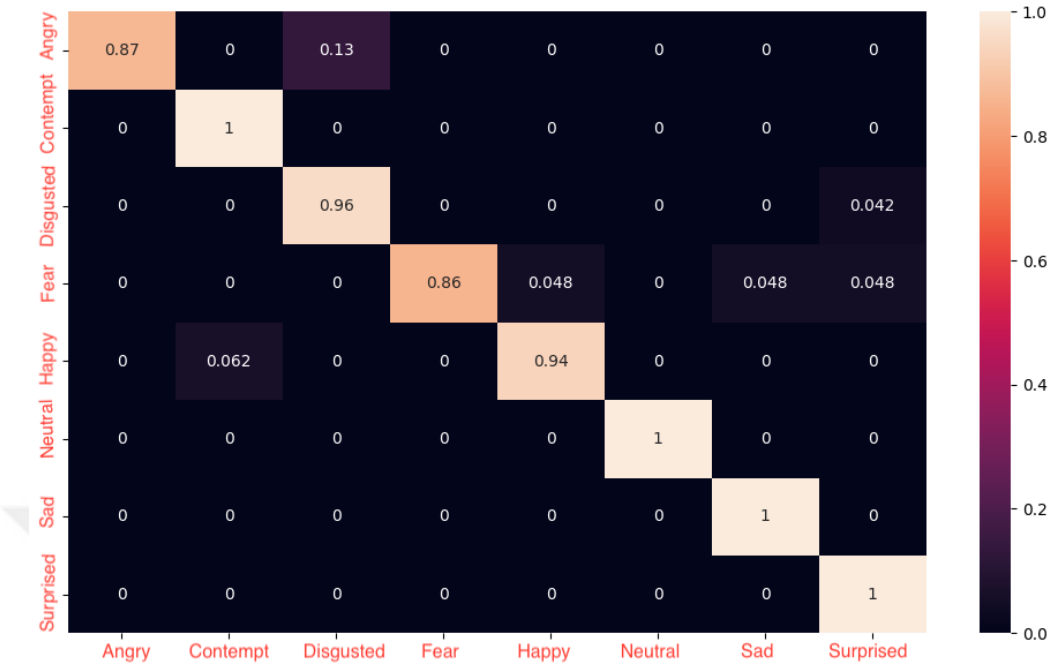

Figure 6.1 Confusion matrix of model tested with MEAD.

Table 6.3 The results of the model with the Ryerson Database.

Emotion Class	Precision	Recall	F1-Score
Angry	1.00	0.73	0.84
Disgusted	0.88	1.00	0.94
Fear	0.89	1.00	0.94
Happy	1.00	0.95	0.98
Sad	1.00	0.76	0.87
Surprised	0.83	1.00	0.94
Calm	0.90	1.00	0.95
Neutral	1.00	0.91	0.95

Ultimately, the model undergoes training and testing using the MELD data set. The prediction outcomes of the MELD are displayed in table 6.5. The confusion matrix findings obtained from testing the model using MELD are displayed in figure 6.3. The comparison of proposed model with other models is presented in table 6.6.

Table 6.4 The comparison of the proposed model with studies in the literature using Ryerson.

Model	**Accuracy (%)**
ResNet18 [53]	57.47
Wav2vec2-PT [54]	84.30± 1.70
CNN-14 [55]	76.58±2.18
CTENet [55]	82.31
MLP [56]	82.75
Audio Based [57]	87.31
Waveforms [58]	90.42
Proposed model	93.00

Figure 6.2 Confusion matrix of model tested with Ryerson.

6.2. CNN Based Model Results

To assess the efficacy of ONN convolutional layers, we have additionally devised an additional model. The additional model consists of CNN layers. It has the same architecture as the ONN based model. The sole distinction lies in the convolution layers of the model, which have been preserved in their original VGG16 form. The ONN layers are substituted

Table 6.5 The results of the model with the MELD.

Emotion Class	Precision	Recall	F1-Score
Angry	0.93	0.78	0.85
Disgust	0.00	0.00	0.00
Fear	0.00	0.00	0.00
Joy	1.00	0.80	0.89
Neutral	0.70	0.98	0.82
Sadness	1.00	0.14	0.25
Surprise	0.69	0.79	0.73

Table 6.6 The comparison of the proposed model with studies in the literature using MELD.

Model	Accuracy (%)
Text-CNN [59]	55.12
text [59]	56.75
CNN [60]	55.02
LSTM [60]	56.44
Text-CNN [61]	59.69
EmoHD [62]	61.27
TA-MERT [63]	64.36
Proposed model	75.00

by convolutional layers. The CNN based model also requires two inputs. The inputs are extracted from the emotion audio-visual data sets. One of the inputs is key frame. The other input is the audio spectrogram.

CNN based model is also trained as the ONN based one is trained. The results of the CNN based model and ONN based one are compared. Therefore, we may see varying effects of the ONN layers on the problem of emotion categorization. The result of CNN model with the data sets are shown in table 6.7. The table also compares CNN based result with the proposed model's results. According to the comparisons, the model based on ONN convolutional

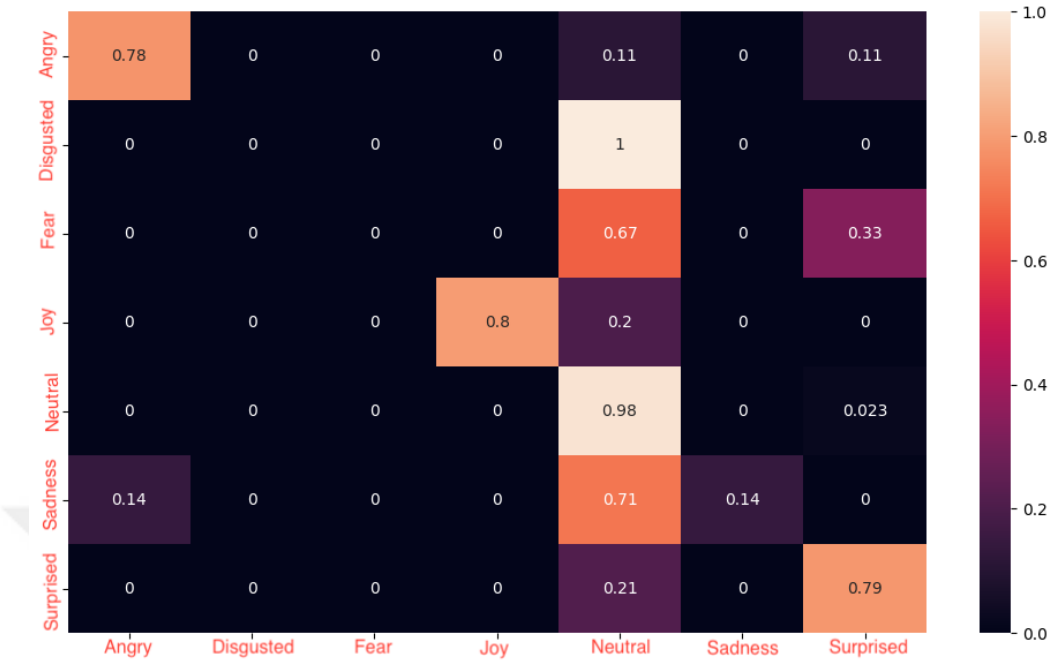

Figure 6.3 Confusion matrix of model tested with MELD.

layers has demonstrated a classification accuracy that is superior to the model employing the original convolutional layers. Hence, it can be asserted that using ONN convolutional layers can enhance accuracy instead of using CNN layers. So, it can be noted that the emotion categorization task demonstrates that ONN yields a highly successful outcome.

Table 6.7 The comparison of the proposed model with CNN based model.

Model	**MEAD Accuracy (%)**	**Ryerson Accuracy (%)**	**MELD Accuracy (%)**
CNN Based Model	50.00	41.00	52.00
Proposed Model	96.00	93.00	75.00

6.3. SNN Based Model Results

As it is mentioned, the main purpose of this study is to propose a new method to classify emotions. ONN and CNN based methodologies are developed and analyzed. In addition to these methodologies, it is aimed at developing another model based on SNN methodologies.

Therefore, another analysis can be undertaken on an alternative and widely-used type of neural network. Within this particular scenario, a research is further conducted using SNN. SNN approaches are studied and evaluated.

Challenges arise while working with the model constructed using SNN. Inaccuracies occur during attempts to mimic the temporal firing of neurons. It employs a distinct computational methodology compared to CNN and ONN. SNN facilitates the execution of transactions in real-time. This study aims to provide information from a certain time period. Hence, the time change factor is inapplicable. When conveying information through spikes, it is possible for some erroneous scenarios to arise. During the course of this investigation, it is important to acknowledge the inclusion of SNN in the operational calculations.

We have developed a rudimentary SNN model that accepts a single visual stimulus as its input. Several forecasts are made. The outcomes do not meet the intended expectations comparing the results of ONN based model. Varying outcomes are anticipated. Given the temporal variability of SNN's structure, it is necessary to modify the model using the LIF theorem. The primary focus of this study has been on ONN approaches. So, we lack sufficient time to construct a very effective model that utilizes SNN approaches to address emotion categorization issues. A novel approach to emotion categorization in future research might involve doing a study utilizing a SNN based approach and a multi-input model.

7. DISCUSSION

A variety of challenges arise while analyzing emotions. Consequently, there might be several gaps in the models. This study introduces a novel emotion categorization model to the existing literature. Significant progress has been made. Nevertheless, doing much research can yield more accurate outcomes. Instead of extracting a single key frame from the video, several time-based pictures may be captured as a time series. Additionally, sound spectrograms can be retrieved from the corresponding time period of these images. Subsequently, research may be conducted to train the model by providing many binary

inputs from a movie. By employing spike neural network time-based learning approaches, this study may be elevated to a higher level. Moreover, the presented model uses VGG16 architecture. VGG16 has a deep but simple and understandable architecture. Small filters are used. It has a modular structure. In this way, changes can be made to the model easily. Image-based classifications have also proven themselves with the parameters they contain. For these reasons, we use the VGG16 architecture as the basis for our study. However, the architectures of other pre-trained models can also be used and results can be obtained. This factor may also be included in future studies.

In addition, we perform an ablation study to assess the performance of the model described in our research. The objective is to comprehend the impact of the model's dual inputs on the resolved issue. An essential characteristic of the model is that it consists of two inputs. To comprehend the impact of these two inputs, an additional model is constructed alongside the existing ones. Our study includes two image inputs in our model. Nevertheless, the model constructed for this ablation investigation exclusively accepts a single input. This entry is also in the format of an image. The layers of the model are identical to the layers of the primary model being presented. Indeed, we select only one of the two branches in our primary model and subsequently refine this branch as an independent model. In this single-input model, we use the data sets we used in the two-input model. The supplementary model has been trained and evaluated. The outcomes derive from the supplementary model for each data set are presented below. The table 7.1 additionally displays the outcomes of our primary model. Hence, it is possible to draw a parallel between the ablation study and the main investigation. Upon examining the findings, it is evident that our two-input model outperforms our single-input model in every dataset. Hence, one might suggest that the two inputs indeed impose a beneficial impact on classification.

8. CONCLUSION

Emotion categorization is the process of identifying human emotions by analyzing text, audio, or pictures. This subject holds significant importance for scholars in fields such as

Table 7.1 The comparison of the proposed model with single input model.

Model	**MEAD Accuracy (%)**	**Ryerson Accuracy (%)**	**MELD Accuracy (%)**
Single Input Model	88.00	80.00	73.00
Proposed Model	96.00	93.00	75.00

finance and customer service. Nevertheless, the process of classifying emotions encounters challenges stemming from cultural disparities among individuals or the utilization of irony in communication. Multiple studies exist on the topic of emotion categorization. Certain research concentrates on a certain form of input for their models. Typically, emotion classification models are constructed using a single input modality, such as sound or visuals. This study introduces innovative design for an operational neural network that supports multiple inputs and many branches. This work attempts to enhance the resolution of emotion categorization challenges by developing a model that incorporates two inputs.

This study specifically examines several neural network methodologies. Initially, we undertake research utilizing the ONN technique. Next, we construct a model with ONN layers. The model necessitates the use of two visual inputs. Every input is simultaneously processed at separate levels. The model consists of two distinct branches. The branches possess an equal number of layers and exhibit identical types of layers. The organization of these branches follows the architecture of VGG16. The model utilizes a multi-stream architecture based on ONN. The conventional convolutional layers utilized in standard CNNs are substituted by ONN convolution layers. The model necessitates audio-visual data sets. Each data point in the data collection is processed. To begin with, the Mel spectrogram is derived from the video. The provided picture is a sound spectrogram that displays the frequencies present in the person's voice. The image is provided and serves as the initial input for the model. Next, a certain frame is chosen from the video. This is an image capturing the facial expression of the individual at that specific instant. The frame serves as an additional input for the model. The model receives both of these inputs concurrently. The model undergoes training and testing using three distinct data sets. Each data set comprises

video recordings featuring one individual. Every video showcases a distinct emotion. Our model has obtained a classification accuracy of 96% from MEAD trials, 93% from Ryerson Emotion Database experiments, and 75% from MELD experiments. We have conducted a comparative analysis of our findings with other empirical investigations utilizing the same data sets. The findings achieved are encouraging and comparable to the current state of the art. Furthermore, we have created a second model that has a comparable structure to the initial model. The only distinction lies in the replacement of ONN layers with CNN convolutional layers. Furthermore, we proceeded to train and evaluate this model. It has been shown that substituting the convolutional layers of commonly used models with ONN layers leads to improved outcomes. Thus, it can be asserted that employing ONN can enhance the performance of models constructed using conventional convolutional layers.

In this study, we have specifically concentrated on and analyzed the impact of various input types on emotion analysis. In addition, we have also seen that ONN operations have a visible impact on emotion classification. We state the opinion that it has the potential to make a valuable contribution to future research. It is expected that employing larger audio-visual data sets would lead to improved performance in models utilizing this strategy. We also want to include more intricate designs, such as Resnets and inception networks using this approach.

REFERENCES

[1] Lakshay Bharadwaj. Sentiment analysis in online product reviews: Mining customer opinions for sentiment classification. *International Journal For Multidisciplinary Research*, 5, **2023**. doi:10.36948/IJFMR.2023.V05I05.6090.

[2] Anuj Kumar and Shashi Shekhar. A survey on sentiment analysis in health care: New opportunities and challenges. pages 621–631, **2023**. doi:10.1007/978-981-99-3608-3_43.

[3] Jyoti Yadav. Sentiment analysis on social media. *Qeios*, **2023**. doi:10.32388/YF9X04.

[4] Jozefien De Leersnyder, Michael Boiger, and Batja Mesquita. Cultural differences in emotions. *Emerging Trends in the Social and Behavioral Sciences*, pages 1–15, **2015**. doi:10.1002/9781118900772.ETRDS0060.

[5] Ruth Filik, Alexandra Țurcan, Dominic Thompson, Nicole Harvey, Harriet Davies, and Amelia Turner. Sarcasm and emoticons: Comprehension and emotional impact. *Quarterly Journal of Experimental Psychology*, 69:2130–2146, **2016**. ISSN 17470226. doi:10.1080/17470218.2015.1106566.

[6] Gianluca Donate, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:974–989, **1999**. ISSN 01628828. doi:10.1109/34.799905.

[7] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, **2008**. ISSN 15540677. doi:10.1561/1500000011.

[8] Adil Baqach and Amal Battou. Text-based sentiment analysis. *Lecture Notes in Networks and Systems*, 637 LNNS:106–121, **2023**. ISSN 23673389. doi:10.1007/978-3-031-26384-2_10.

[9] Nur Alia Syahirah Badrulhisham and Nur Nabilah Abu Mangshor. Emotion recognition using convolutional neural network (cnn). *Journal of Physics: Conference Series*, 1962, **2021**. ISSN 17426596. doi:10.1088/1742-6596/1962/1/012040.

[10] Attention-based modeling for emotion detection and classification in textual conversations — request pdf.

[11] K N ArulJothi, Sivaraj Irusappan, Gautami Amarnath, Swathine Chandrasekaran, Sruthi Abirhami B K, MK Harishankar, Janitri V Babu, and Devi A. Emotion detection through audio using machine learning. *IJSR - International Journal of Scientific Research*, Volume 8 Issue 1:63–65, **2019**. ISSN 2249 - 555X. doi:10.36106/IJSR.

[12] Swayam Badhe and Sangita Chaudhari. Deep learning based facial emotion recognition. *ITM Web of Conferences*, 44:03058, **2022**. doi:10.1051/ITMCONF/20224403058.

[13] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, **2020**. ISSN 21693536. doi:10.1109/ACCESS.2020.3026823.

[14] Giovanni Cerulli. Artificial neural networks. pages 269–322, **2023**. doi:10.1007/978-3-031-41337-7_6.

[15] Shreyas D. K., Srivatsa N. Joshi, Vishwas H. Kumar, Vishaka Venkataramanan, and Kaliprasad C. S. A review on neural networks and its applications. *Journal of Computer Technology Applications*, **2023**. doi:10.37591/JOCTA.V14I2.1062.

[16] Eman Jawad. The deep neural network-a review. *IJRDO -JOURNAL OF MATHEMATICS*, 9:1–5, **2023**. doi:10.53555/M.V9I9.5842.

[17] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark.

[18] Hiroyoshi Todo, Tianqi Chen, Jiazhen Ye, Bin Li, Yuki Todo, and Zheng Tang. Single-layer perceptron artificial visual system for orientation detection. *Frontiers in Neuroscience*, 17, **2023**. ISSN 1662453X. doi:10.3389/FNINS.2023.1229275.

[19] Gordon Lightbody and George W. Irwin. Multi-layer perceptron based modelling of nonlinear systems. *Fuzzy Sets and Systems*, 79:93–112, **1996**. ISSN 0165-0114. doi:10.1016/0165-0114(95)00293-6.

[20] Shao Yu Yin, Yu Huang, Tien Yu Chang, Shih Fang Chang, and Vincent S. Tseng. Continual learning with attentive recurrent neural networks for temporal data classification. *Neural Networks*, 158:171–187, **2023**. ISSN 18792782. doi:10.1016/J.NEUNET.2022.10.031.

[21] Ga-Ae Ryu, Tserenpurev Chuluunsaikhan, Aziz Nasridinov, HyungChul Rah, and Kwan-Hee Yoo. Sce-lstm: Sparse critical event-driven lstm model with selective memorization for agricultural time-series prediction. *Agriculture*, 13:2044, **2023**. doi:10.3390/AGRICULTURE13112044.

[22] Anirudha Ghosh, Abu Sufian, Farhana Sultana, Amlan Chakrabarti, and Debashis De. Fundamental concepts of convolutional neural network. *Intelligent Systems Reference Library*, 172:519–567, **2019**. ISSN 18684408. doi:10.1007/978-3-030-32644-9_36.

[23] Shaharyar Ahmed Khan Tareen and Filza Khan Tareen. Convolutional neural networks for beginners. *SSRN Electronic Journal*, **2023**. doi:10.2139/SSRN.4566310.

[24] Andreas Waldis, Luca Mazzola, and Michael Kaufmann. Concept extraction with convolutional neural networks. *DATA 2018 - Proceedings of the 7th International Conference on Data Science, Technology and Applications*, pages 118–129, **2018**. doi:10.5220/0006901201180129.

[25] Shruti R. Kulkarni, Anakha V. Babu, and Bipin Rajendran. Spiking neural networks - algorithms, hardware implementations and applications. *Midwest Symposium on Circuits and Systems*, 2017-August:426–431, **2017**. ISSN 15483746. doi:10.1109/MWSCAS.2017.8052951.

[26] M. Prezioso, M. R. Mahmoodi, F. Merrikh Bayat, H. Nili, H. Kim, A. Vincent, and D. B. Strukov. Spike-timing-dependent plasticity learning of coincidence detection with passively integrated memristive circuits. *Nature Communications 2018 9:1*, 9:1–8, **2018**. ISSN 2041-1723. doi:10.1038/s41467-018-07757-y.

[27] Kashu Yamazaki, Viet Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. Spiking neural networks and their applications: A review. *Brain Sciences*, 12, **2022**. ISSN 20763425. doi:10.3390/BRAINSCI12070863.

[28] Doron Tal and Eric L. Schwartz. Computing with the leaky integrate-and-fire neuron: Logarithmic computation and multiplication. *Neural Computation*, 9:305–318, **1997**. ISSN 08997667. doi:10.1162/NECO.1997.9.2.305.

[29] Serkan Kiranyaz, Turker Ince, Alexandros Iosifidis, and Moncef Gabbouj. Operational neural networks. *Neural Computing and Applications*, 32:6645–6668, **2020**. ISSN 14333058. doi:10.1007/S00521-020-04780-3/TABLES/6.

[30] Christopher Edmond. Classification performance for credit scoring using neural network. *International Journal of Emerging Trends in Engineering Research*, 8:1592–1599, **2020**. doi:10.30534/IJETER/2020/19852020.

[31] Junaid Malik, Serkan Kiranyaz, and Moncef Gabbouj. Fastonn – python based open-source gpu implementation for operational neural networks. **2020**.

[32] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained

models: Past, present and future. *AI Open*, 2:225–250, **2021**. ISSN 26666510. doi:10.1016/j.aiopen.2021.08.002.

[33] Yankai Lin, Ning Ding, Zhiyuan Liu, and Maosong Sun. Pre-trained models for representation learning. *Representation Learning for Natural Language Processing*, pages 127–167, **2023**. doi:10.1007/978-981-99-1600-9_5.

[34] Jiazhi Liang. Image classification based on resnet. *Journal of Physics: Conference Series*, 1634, **2020**. ISSN 17426596. doi:10.1088/1742-6596/1634/1/012110.

[35] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. **2023**.

[36] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, and Marco Andreetto. Mobilenets: Efficient convolutional neural networks for mobile vision applications.

[37] Prerepa Gayathri, Aiswarya Dhavileswarapu, Sufyan Ibrahim, Rahul Paul, and Reena Gupta. Exploring the potential of vgg-16 architecture for accurate brain tumor detection using deep learning. *Journal of Computers, Mechanical and Management*, 2, **2023**. doi:10.57159/GADL.JCMM.2.2.23056.

[38] Alireza Rahimzadeganasl and Elif Sertel. Automatic building detection based on cie luv color space using very high resolution pleiades images. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. **2017**. doi:10.1109/SIU.2017.7960711.

[39] Andreas Koschan and Mongi Abidi. *Color Spaces and Color Distances*, pages 37–70. **2007**. ISBN 9780470147085. doi:10.1002/9780470230367.ch3.

[40] Youn Jin Kim, M Ronnier Luo, Peter Rhodes, Vien Cheung, Stephen Westland, Seungsin Lee, Youngshin Kwak, Dusik Park, and Changyeong Kim. Measurement of perceived brightness and contrast sensitivity for outdoor condition.

[41] Fabian Immanuel Ijpelaar. Entropy in physics: An overview of definitions and applications in quantum mechanics. **2021**.

[42] Yanping Zhao and Xiaolai Zhou. K-means clustering algorithm and its improvement research. *Journal of Physics: Conference Series*, 1873, **2021**. ISSN 17426596. doi:10.1088/1742-6596/1873/1/012074.

[43] Spatial filters - laplacian/laplacian of gaussian.

[44] Boyang Zhang, Jared Leitner, and Sam Thornton. Audio recognition using mel spectrograms and convolution neural networks.

[45] Md Afzal Hossan, Sheeraz Memon, and Mark A. Gregory. A novel approach for mfcc feature extraction. *4th International Conference on Signal Processing and Communication Systems, ICSPCS'2010 - Proceedings*, **2010**. doi:10.1109/ICSPCS.2010.5709752.

[46] Serkan Kiranyaz, Junaid Malik, Habib Ben Abdallah, Turker Ince, Alexandros Iosifidis, and Moncef Gabbouj. Self-organized operational neural networks with generative neurons. *Neural Networks*, 140:294–308, **2020**. ISSN 18792782. doi:10.1016/j.neunet.2021.02.028.

[47] Junaid Malik, Serkan Kiranyaz, and Moncef Gabbouj. Self-organized operational neural networks for severe image restoration problems. *Neural Networks*, 135:201–211, **2021**. ISSN 0893-6080. doi:10.1016/J.NEUNET.2020.12.014.

[48] Luis A. Camuñas-Mesa, Bernabé Linares-Barranco, and Teresa Serrano-Gotarredona. Neuromorphic spiking neural networks and their memristor-cmos hardware implementations. *Materials*, 12, **2019**. ISSN 19961944. doi:10.3390/MA12172745.

[49] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12366 LNCS:700–717, **2020**. ISSN 16113349. doi:10.1007/978-3-030-58589-1_42/COVER.

[50] The ryerson audio-visual database of emotional speech and song (ravdess). doi:10.5281/ZENODO.1188976.

[51] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 527–536, **2018**. doi:10.18653/v1/p19-1050.

[52] Samir Sadok, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda, and Renaud Séguier. A multimodal dynamical variational autoencoder for audiovisual speech representation learning. pages 3–3, **2023**. doi:10.1145/3552487.3556435.

[53] Sumon Kumar Hazra, Romana Rahman Ema, Syed Md Galib, Shalauddin Kabir, and Nasim Adnan. Emotion recognition of human speech using deep learning method and mfcc features. *Radioelectronic and Computer Systems*, 2022:161–172, **2022**. ISSN 26632012. doi:10.32620/REKS.2022.4.13.

[54] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1:551–555, **2021**. ISSN 19909772. doi:10.21437/Interspeech.2021-703.

[55] Rizwan Ullah, Muhammad Asif, Wahab Ali Shah, Fakhar Anjam, Ibrar Ullah, Tahir Khurshaid, Lunchakorn Wuttisittikulkij, Shashi Shah, Syed Mansoor Ali, and Mohammad Alibakhshikenari. Speech emotion recognition using

convolution neural networks and multi-head convolutional transformer. *Sensors 2023, Vol. 23, Page 6212*, 23:6212, **2023**. ISSN 1424-8220. doi:10.3390/S23136212.

[56] Surbhi Khurana, Amita Dev, and Poonam Bansal. Robustness evaluation of multi-layer perceptron based speech emotion recognition model for hindi language. page 030012. **2023**. doi:10.1063/5.0177794.

[57] Selma Ozaydin. Emotional speech recognition based on cnn. *INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING TECHNOLOGY*, 14:60–64, **2023**. doi:10.17605/OSF.IO/2CZ7A.

[58] Zeynep Kilimci, Ulku Bayraktar, and Ayhan Kucukmanisa. Evaluating raw waveforms with deep learning frameworks for speech emotion recognition. **2023**.

[59] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 527–536, **2018**. doi:10.18653/v1/p19-1050.

[60] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, **2019**. ISSN 21693536. doi:10.1109/ACCESS.2019.2929050.

[61] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 7042–7052, **2021**. doi:10.18653/v1/2021.acl-long.547.

[62] Aditi Sharma, Kapil Sharma, and Akshi Kumar. Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. *Neural Computing and Applications*, 35:1–14, **2022**. doi:10.1007/s00521-022-06913-2.

[63] Fida Mohammad, Mukhtaj Khan, Safdar Nawaz Khan Marwat, Naveed Jan, Neelam Gohar, Muhammad Bilal, and Amal Al-Rasheed. Text augmentation-based model for emotion recognition using transformers. *Computers Materials Continua*, 76:3523–3547, **2023**. doi:10.32604/cmc.2023.040202.