DEVELOPMENT OF MULTIMODAL PERSONALITY PREDICTION MODEL USING PERSONALITY TRAITS

A DISSERTATION

Submitted in the partial fulfilment of the requirements
OF
Master of Technology
In
INFORMATION SYSTEMS

Submitted by:
ANISHA PATEL
2K21/ISY/04
Under the supervision of

PROF. DINESH K. VISHWAKARMA

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Bawana Road, Delhi – 110042

MAY 2023

CANDIDATE'S DECLARATION

Anisha Patel (2K21/ISY/08), Student of MTech (Information Systems), hereby declare that the project dissertation titled "**Development of Multimodal Personality Prediction Model using Personality Traits**", which is submitted by me to the department of Information Technology, Delhi Technological University, Delhi in fulfilment of the requirement for the award of degree of MASTER OF TECHNOLOGY is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associate ship, Fellowship or other similar title or recognition.

Place: Delhi

Date: 31ST May, 2023

Anisha Patel

(2K21/ISY/04)

CERTIFICATE

This is to certify that the project dissertation titled "**Development of Multimodal Personality Prediction Model using Personality Traits**" which is submitted by Anisha Patel(2K21/ISY/04), Department of Information Technology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Technology is a record of the project work carried out by the student under my supervision.

Place: Delhi

Date:31st May,2023

Dr. Dinesh K. Vishwakarma

SUPERVISOR

ACKNOWLEDGEMENT

I am very thankful to Dr. Dinesh K. Vishwakarma (Professor, Department of Information Technology) and all the faculty members of the Department of Information Technology at DTU. They all provided us with immense support and guidance for the project. I would also like to express my gratitude to the University for providing us with the laboratories, infrastructure, testing facilities and environment which allowed us to work without any obstructions. I would also like to appreciate the support provided to us by our lab assistants, seniors and our peer group who aided us with all the knowledge they had regarding various topics.

Anisha Patel

Roll No. 2K21/ISY/04

ABSTARCT

It might be challenging to predict someone's personality in both the workplace and in daily life. There are several variables that can affect personality prediction, and these variables can change from person to person. Personality reflects an individual’s behaviour, thought process , life choices, mental health , emotions, social character. Variou deep learning model has been used in this project for multi modal personality prediction. VGGish convolutional networks (VGGish CNN), Resnet50, InceptionV3, Xception, Convnext, InceptionResnet have been used to extract facial and ambient features from the video, 2D convolutional neural network and Alexnet, have been to extract audio features. For the final prediction the extracted feature is given as a input to a fully connected layer followed by sigmoid activation function.

Keywords—Big five traits, convolutional neural network, multimodal fusion, personality prediction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

The personality prediction using videos is a significant problem in the field of personality computing. Personality prediction from multimodal data has become an emerging research area in computer vision as a result of recent advances in machine learning. There are a plethora of potential application where personality plays an important role and It is possible to use a system that can recognise people's perceived personality traits with accuracy. The model used to describe human personality most commonly is the Big Five personality characteristics.[5]

Psychologists and behavioural scientists have long been interested in comprehending and predicting human personality traits. Applications in many fields, including psychology, human-computer interaction, and customised recommendation systems, are made possible by accurate personality evaluation, which offers insightful knowledge into people's behaviours, preferences, and inclinations. Self-report questionnaires and interviews, which can be deceptive, time-consuming, and biased, have historically been the foundations of personality evaluation. Recent advances in machine learning techniques have created new opportunities for personality prediction. The method of predicting a person's personality traits utilising different input modalities, such as text, images, audio, and video, is known as multimodal personality prediction. By using a range of information sources, multimodal approaches strive to provide a more complete and precise prediction of a person's personality.

Recent advances in machine learning techniques have created new opportunities for personality prediction. These models are excellent at learning hierarchical representations automatically from unstructured data, which eliminates the need for manual feature engineering. Researchers can use deep learning to build more reliable and precise personality prediction models by taking advantage of the intricate interactions found in multimodal data.

This study seeks to determine how well deep learning models predict the Big Five personality traits from multimodal data. We seek to create a thorough and nuanced

understanding of people's personality profiles by merging many modalities, such as text, photos, and voice. Additionally, in order to determine the best strategy for personality prediction, this study compares the performance of various deep learning architectures created expressly for multimodal data, such as convolutional neural networks (CNNs), and other models.

The early 1980s saw the development of the Five Factor Model (FFM), often referred to as the OCEAN model and the Big Five model[2], as a consequence of extensive research into many psychological theories. This model is based on statistical analyses done on data from personality surveys where people are described using particular terms, giving a succinct and accurate description of their general character or personality. These five traits are stored in dataset represented with a value within a range [0,1].

- **Open to Experience**: This character quality describes a person's curiosity, openness, and desire to try new things. Those with great openness tend to be creative, inquisitive, and tolerant. They like to investigate novel thoughts, ideas, and experiences. Those with low openness may be more conservative, cautious, and accustomed to innovation.
- **Conscientiousness** This characteristic is used to indicate a person's diligence and attention. It's a characteristic that describes how organised and effective someone is. The degree of organisation, responsibility, and self-discipline in a person is what defines conscientiousness. People with a high level of conscientiousness are trustworthy, diligent, and reliable. They have a strong sense of obligation, establish specific objectives, and strive assiduously to achieve them. Conversely, those with low conscientiousness may be more impulsive, disorganised, and uninterested in long-term planning.
- **Extraversion**: The ability to engage with others is the characteristic that best characterises a candidate's social abilities. Extraversion describes how much a person seeks out social engagement, mental stimulation, and appreciates being with other people. Extraverts are frequently characterised as gregarious, chatty, and enthusiastic. They thrive in social settings and get their energy from being around other people. On the other side, introverts are more reserved, prefer quieter settings, and occasionally require some alone time to recharge.

- **Agreeableness:** It is a trait that evaluates a person's behaviour based on their generosity, compassion, cooperation, and capacity for interpersonal adjustment. A person's ability to get along with others and their general attitude towards kindness, cooperation, and empathy are both influenced by their agreeableness. Friendly, sympathetic, and sensitive of others' needs are all characteristics of highly agreeable people. They are frequently referred to as friendly and nurturing. People with low agreeableness may be more forceful, competitive, and less concerned with the thoughts and feelings of others.
- **Neuroticism:** This characteristic often depicts a person as having strong expressive power and mood swings. Neuroticism is the degree of emotional stability and inclination to experience negative emotions including sadness, anxiety, and mood swings. Greater emotional reactivity, concern, and stress are all correlated with high levels of neuroticism. Less neurotic people are often more emotionally secure, resilient, and less prone to stressful emotional occurrences.

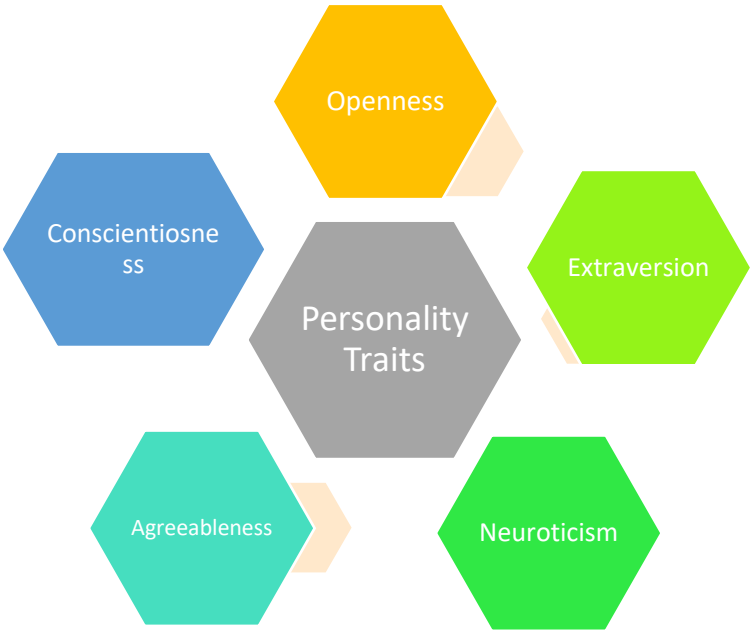

Fig1. 1 The Big Five Model

1.1 MOTIVATION

The idea behind multimodal personality production comes from the knowledge that personality is a complicated, multidimensional feature that is impacted by a variety of elements, including both audio and visual clues. The personality of a person can be better understood by integrating data from many modalities, such as speech, body language, and facial expressions.

A single modality is unable to properly express the complexity of personality. Using a single modality to predict personality might lead to uncertainty and noise. We can reduce the limitations and uncertainties connected to any one modality by merging many modalities. The accuracy of personality prediction models may be enhanced by incorporating data from many modalities. Each modality provides distinctive insights into a person's mental state, and by combining them, we may take use of their respective strengths to create a more accurate model.

Aspects of human behaviour and communication are captured differently by various modalities. For instance, though body language and facial expressions give non-verbal indicators, speech may convey linguistic patterns, voice tones, and verbal emotions. We may get a more complete picture of a person's personality by taking into consideration a variety of modalities and including both verbal and non-verbal indicators.

The overall objective of multimodal personality prediction is to increase the precision, richness, of personality prediction models by using the capabilities of several modalities. We can find out more about human behaviour, communication, and individual variations by taking into account both audio and visual inputs. This unlocks up possibilities for applications that are necessary for a detailed comprehension of human behaviour.

1.2 Organization of Dissertation:

In the subsequent section, we will take a gander at the point overall and how it became so significant, featuring how helpful it might have been. The fundamentals of the deep neural network model will be discussed in detail in the second chapter. We will likewise talk about related research business related to this review.

The architecture proposed convolutional neural network model for predicting human personality and the approach of the convolutional neural network model is discussed in the third chapter.

The experimental technique will be covered in chapter 4. We'll further look at the project's data analysis and dataset.

In chapter 5, we will cover performance metrics in this chapter and discuss how to compute them. The sixth chapter of the thesis is focused on a conclusion and includes further research options.

1.3 DATASETS:

1. **Chalearn-17: First Impressions V2 dataset** has been used for this project, the dataset contains mp4 clips and the ground truth annotations. It has 10,000 video clips each of 15 seconds with a person talking to a camera. Different age, gender, nationality, and ethnic groupings are represented in the videos. Moreover the database includes several unique cases, such as on In some database videos, individuals might be heard speaking sign language or in situations where a person is seated in front of the camera without making a sound or moving. Each video has six labels, each with a value between 0 and 1. Extraversion, agreeableness, conscientiousness, neuroticism, and openness are the five major personality qualities that are covered by five of them. Amazon Mechanical Turk (AMT) employees have gathered these YouTube videos and labelled them with the Big Five characteristics. The dataset is pre-processed then this pre-processed dataset is used to classify or predict user personality[15].

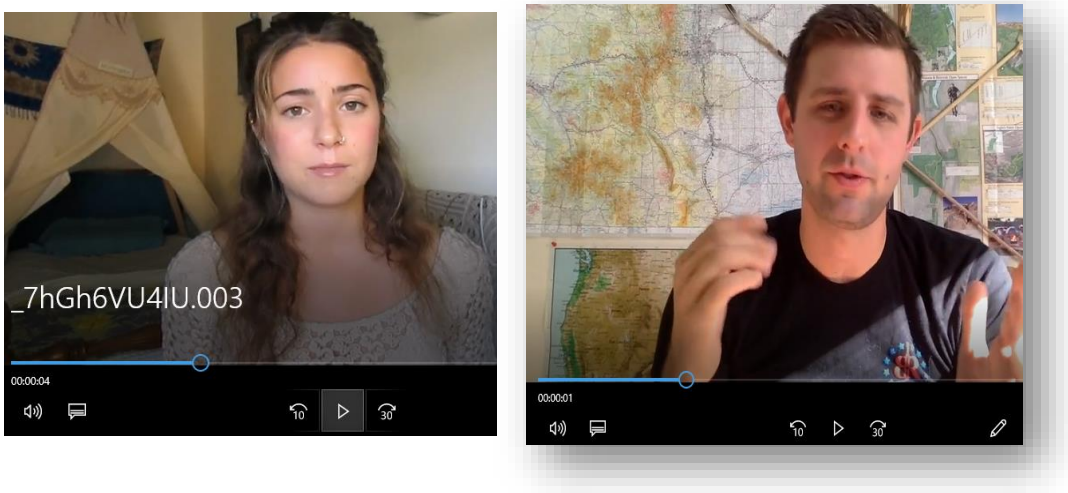

Fig1.2. video examples from the ChaLearn Looking At People competition dataset

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 LITERATURE SURVEY

The author [1] focuses on a fascinating area of research: predicting personality traits using a deep learning-based approach. The study explores the use of multiple data modalities, including visual, audio, and text, to gain insights into an individual's behavior, mental health, emotions, life choices, social nature, and thought patterns. The researchers developed a system that extracts features from facial and ambient cues, audio signals, and text using advanced convolutional neural network models. These extracted features are then combined using fusion techniques such as concatenation and attention mechanisms. The system's performance was evaluated using a dataset called Chalearn-17, and the results indicate that combining features from different modalities can achieve comparable results to traditional averaging methods. Interestingly, the study also found that even a small number of images can provide sufficient information for accurate personality prediction. Overall, this research offers valuable insights into the potential of multi-modal approaches for automatic personality prediction, with implications for fields such as cyber forensics, personalized services, and recommender systems.

The authors in [3] has designed a model for automatic personality prediction To extract features from the video they used random decision forest to do regression. In the first track of the second round of the Joint Contest on Multimedia Challenges Beyond Visual Analysis at ICPR 2016.

The researchers in [4] has proposed Deep Bimodal regression (DBR) framework for personality prediction. This framework has 3 steps: deep regression of the visual modality, second part is deep regression of audio and the last step is fusion of two modalities..

They extracted rich information of audio and visual modalities by proposing Deep Bimodal Regression framework (DBR). In visual modality of DBR framework they modified CNN architecture by adding an additional layer assembled with max pooling and average layer is added between last convolution layer and fully connected layer, they used CNN with the VGG-16 architecture, for audio modality they extracted

MFCC and logfbank features from the audio, used a model consists of fully connected layer followed by sigmoid layer to train linear regressor for audio modality. Their DBR model won the 1st place in the competition organised by ECCV Chalearn LAP 2016 [4]

In this paper the author explored several modalities such as transcribed speech, voice, head pose and gaze, facial appearance, body gesture and action units. They have evaluated the big five traits on two data sets namely First Impression dataset (FID) and Self-presentation and Induced Behaviour Archive for Personality Analysis (SIAP). For facial appearance modality author used two different architecture 3D ResNext-101 and Convolutional Neural Network-Gated Recurrent Unit combination (CNN-GNU). Spatial relations can be modelled by CNN and temporal relations using recurrent network. Feature extraction of facial action units, head pose and gaze is done by Openface and are modelled by using two different architecture Long- and Short-term Time-series Network (LSTNet) and Recurrent Convolutional Neural Networks (RCNN), for voice features are modelled by LSTnet architecture and they applied BERT embedding & used LSTnet architecture for transcribed speech modality. For the assessment the performance they combined the different modalities with five different method and found that late fusion using LSVR model has performed best on FID datasets. Further they analysed and compare their performance with MAEs of different models on FID[3]

In this paper they introduced several models that were trained on chalearn First impression dataset which contains YouTube High-Definition videos where people are speaking in English in front of the camera, to predict big five personality traits which can help in determining in if the candidate is suitable for the recommendation of job interview. Their model won the 3rd place in the competition organised by ECCV Chalearn LAP 2016. They explored different modalities like visual, audio, language, audio-visual and combination of language and audio-visual and observed that best performance can be achieved by fusing all the modalities.[4]

The authors in this paper used three sets of features: audio, visual and facial they used random decision forest with 5000 trees to do regression from the extracted features. They investigated the use of components of the visual other than the face to emphasise gestures and body language articulations using weighted motion energy images(wMEI).

The movements and activity of facial gestures are captured with the facial landmark features. Their model won the 4th place in the competition organised by ECCV Chalearn LAP 2016 with 10-fold cross validation accuracy 90.19% and test accuracy 0.898[5].

To reduce the time consumption on candidate selection during interview process the author in this paper comes up with the system to predict candidate scores of five big personality traits to make a final decision whether the candidate should be called for an interview or not. In their proposed system they used speech, images and set of words (transcribed speech) from the video. They extracted features from each of the modality and trained them using open-source software TenserFlow. To predicts personality traits they applied multilayer perceptron with two hidden layers and achieved 89% accuracy [7].

They created an AVI using a tensor flow-based semisupervised DL model to accurately auto perceive an interviewee's true personality based on 120 real-world samples of job applicants. They used professional participants, which may limit the generalizability of the results. [8]

Some datasets that has been used by authors

1. Personality Analysis Self-presentation and Induced Behavior Archive (SIAP): In order to elicit indications of personality characteristics, SIAP contains recordings for speaking (question-answering) and watching video clips. The total number of sessions is 900, divided between 180 sessions for the interview portion (60 participants x 3 questions), and 180 sessions for the induction of each characteristic (60 participants , 3 inducing videos for each trait , 5 traits).[9]
2. myPersonality dataset: A well-known Facebook application called myPersonality dataset was developed in 2007 and allowed users to take actual psychometric evaluation tests and view the results right away. One of the biggest study libraries in the history of social science is the consequence of almost 40% of respondents choosing to contribute data from their Facebook profile. Each Big five model component has about 5000 positive and 5000 negative statuses attached to it in the MyPersonality dataset used for this study.[14]

TABLE 2.1: Analysis of personality prediction based on previous paper

Ref.	Models	Dataset +Accuracy	limitations
Dersu Giritlio ˘glu et. al (2020)[9]	3D ResNext-101 Convolutional Neural Network-Gated Recurrent Unit combination (CNN-GNU), LSTNet, RCNN	Self-presentation and Induced Behavior Archive for Personality Analysis (SIAP) with MAE- 0.085 and ChaLearn LAP First Impressions Dataset	Compared to Neuroticism, Openness, and Extraversion, Agreeableness and Conscientiousn ess have observed variations that are substantially smaller.
Rob van Lier et al. (2018)[13]	Residual network	ECCV 2016 First Impression Dataset with accuracy 0.91	The interview notes and the disparities between the top and lowest levels of each feature were clearly visible.
Berkay Aydin et al. (2016)[3]	Random Decision Tree	First Impression dataset with 10 fold cross validation accuracy 0.90197 and test accuracy 0.898	Extraction of audio features were not done effectively because of

			incompatibility of toolbox they used
WMKS Ilmini et al. (2022)[7]	Grad-CAM, Guided Backpropagation, and Guided Grad-CAM techniques.	First Impression V2 dataset with accuracy more than 90% for all the traits except Neuroticism	The output of the visualisation approaches fails to identify the facial areas that mostly contribute to the corresponding personality score for the lowest rated data.
Jelena Gorbova et al. (2017)[16]	Perceptron neural network	First Impression dataset with 89% accuracy on 6 labels	
HUNG-YUE SUEN et al.	CNN	Above 90%	Their dataset contains only 120 participants moreover they used professionals as participants
Menasha Thilakaratne et al.[14]	SVM	myPersonality dataset	
K Rakesh Bharadwaj et al. [9]	Random Forest Regressor and Classfiers	myPersonality dataset with 64.25% accuracy	Accuracy is very low

MICHAEL M. TADESSE et al. [10]	XGboost algorithm	myPersonality dataset with 74.2%	The accuracy seemed to be low because of small number of users from the dataset.

2.2 TAXONOMY

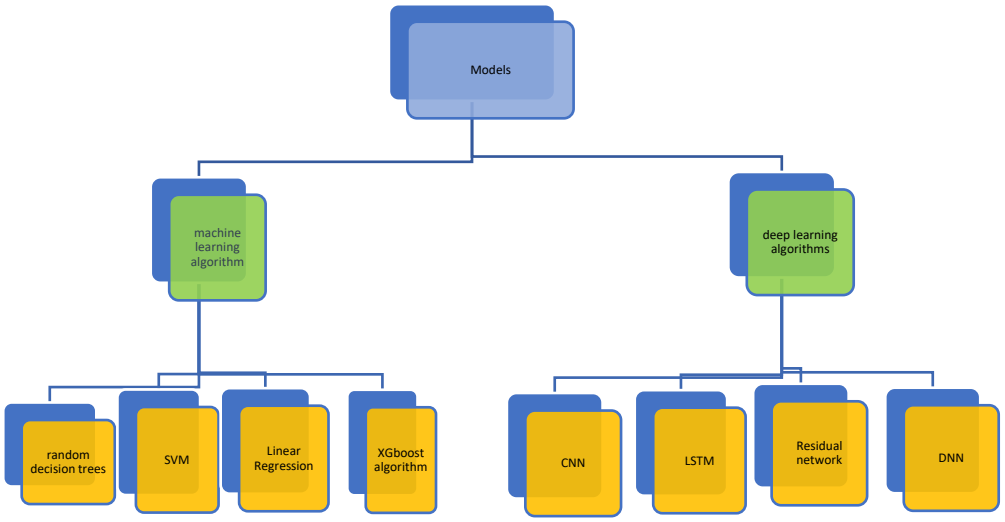

Fig2.1. Models utilized in predicting personality traits

2.2.1 MACHINE LEARNING MODELS

- **Random forest:** Random Forest functions by producing a lot of decision trees. Based on the maximum values in the tree, a final choice is made. To increase the accuracy of its predictions, it employs the average of its results. It is supervised and has both classification and regression applications. In this method, the decision tree is constructed around the k number of random points[9]
- **Support Vector Machine**: To estimate the membership degree of the individuals for each personality trait [9] used linear regression and SVM for regression.

- **Extreme Gradient Boosting algorithm**: The XGBoost method is a decision tree-based machine 3a 10-fold cross-validation with 10 iterations to train our data. A single fold was utilised for testing each time, while the other nine folds were used for training[10]

2.2.2 DEEP LEARNING

DL is a branch of Machine Learning that focuses on training of ANNs with multiple layers, it takes a large amount of data to learn and make predictions. It is inspired by human brain. DL algorithms are made to automatically recognise and extract features from unstructured input, such photos, texts and audios.

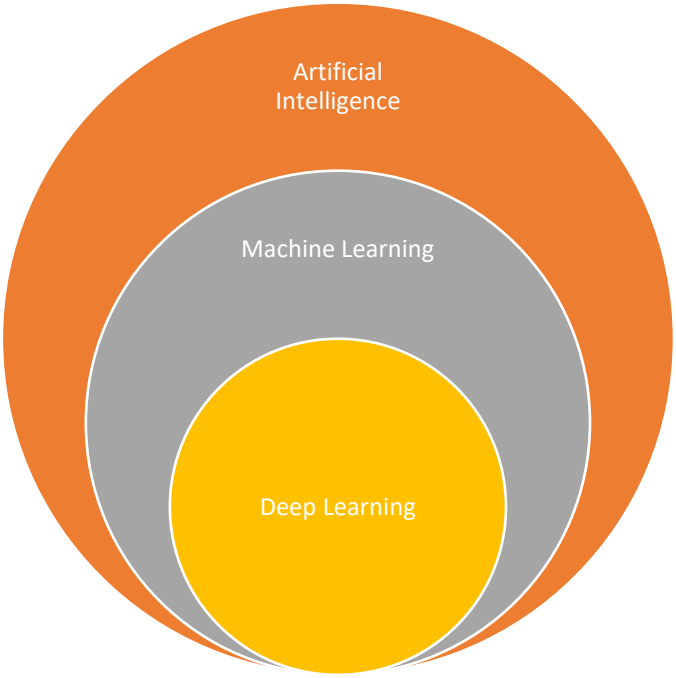

Figure 2.2: Artificial Intelligence v. Machine Learning v. Deep Learning

The main job of a neural network is to take in a variety of inputs, analyse them using sophisticated computations, and output the outcomes to address real-world issues like categorization.

Convolutional Neural Network (CNN)

Convolutional Neural Networks (CNNs) are powerful deep learning architectures that can learn directly from data. While CNNs are widely known for their ability to identify patterns in images and recognize objects, their applicability extends beyond visual data. CNNs can be effectively used for classifying various types of non-image data, such as audio, time series, and signal data. The architecture of a CNN consists of multiple layers, including convolutional layers, pooling layers, and fully connected layers. These layers work together to process and extract relevant features from the input data, enabling the network to differentiate between different inputs. One of the advantages of CNNs is that they require less preprocessing compared to other classification algorithms, making them a versatile and efficient choice for various machine learning tasks.

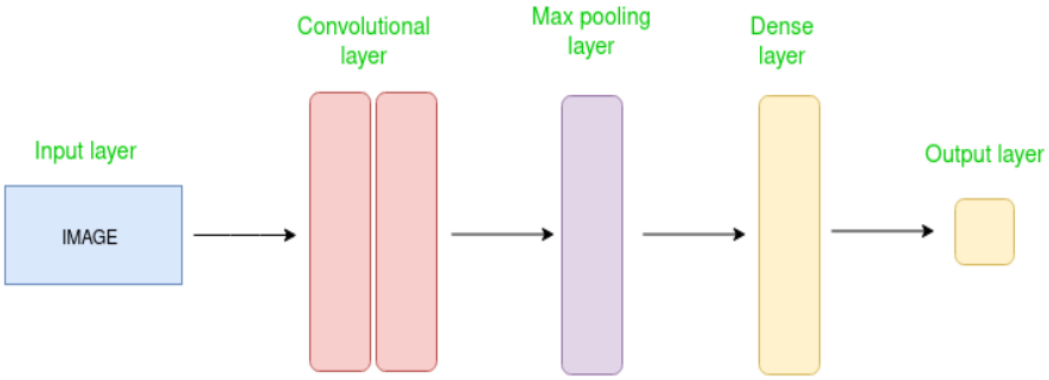

Fig 2.3 Architecture of CNN

Using multi-modal data including images, text, audio, and video, CNNs are capable of carrying out challenging tasks. ConvNet designs include LeNet, ResNet, VGGNet, AlexNet, and GoogLeNet, among others.

Kernel

The kernel in a CNN is only a filter used to extract features from the images.

Layers used to build CNN

When given inputs like image, voices, or sounds, convolutional neural networks perform better than traditional neural networks.

- Convolutional Layers
- Pooling Layers
- Activation Layers
- Fully Connected Layers
- Dropout Layers
- Batch Normalization Layers
- Flatten Layers

Convolutional Layers: These layers use convolution operations to add a number of learnable filters to the input data. Each filter applies element-wise multiplications and summations to the input to extract local patterns or characteristics. The presence of these features at various spatial places is represented by the feature maps that are produced.

Pooling Layers The spatial dimensions of the feature maps can be decreased while still retaining the most important information by merging layers. The pooling techniques max pooling and average pooling are often used by CNNs. Max pooling selects the maximum value that is feasible inside a certain region, whereas average pooling calculates the average value. Pooling lowers the computational complexity and increases the network's resistance to spatial changes.

Activation Layers: Activation layers in neural networks play a crucial role in introducing non-linearity to the model. They take the output from the preceding layer and apply a non-linear activation function to each element. One commonly used activation function in CNNs is the Rectified Linear Unit (ReLU). ReLU helps the network learn complex relationships and patterns in the data by allowing positive values to pass through unchanged while setting negative values to zero. This non-linear activation function enhances the network's ability to capture and represent intricate correlations in the data, making it a popular choice in CNN architectures.

Fully Connected Layers: Also referred to as dense layers, fully connected layers are the standard neural network levels in which every neuron is coupled to every other neuron in every layer below it. These layers carry out sophisticated classification and reasoning

based on the features that were retrieved from lower layers. The CNN architecture often uses fully connected layers at the very end.

Dropout Layers: During training, dropout layers randomly deactivate a portion of the neurons, preventing overfitting and promoting the network to acquire more reliable representations. By making the network learn redundant representations, it lessens the dependence of neurons on one another.

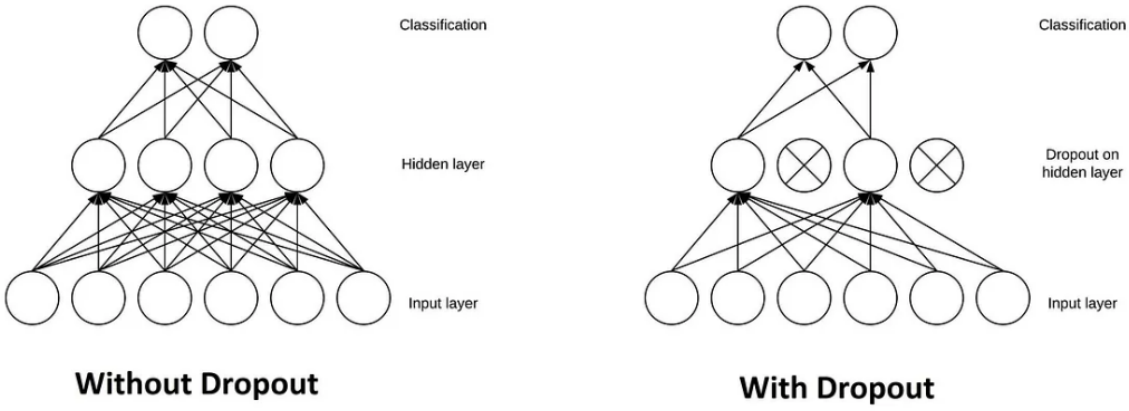

Fig 2.4 Dropout layer in the neural network

Batch Normalisation Layers: By modifying and scaling the activations, batch normalisation layers normalise the input to a layer. This method minimises the internal covariate shift problem, which helps to stabilise and accelerate the training process.

Flatten Layers: The multi-dimensional feature maps are transformed into a one-dimensional vector using the flatten layers command. This shift from the spatial representation to a format appropriate for conventional neural network layers is frequently made before the fully linked layers.

Pooling

After convolutional layers, pooling layers are frequently applied to lower the spatial dimensions of the feature maps while maintaining the most important data. The feature maps can be down sampled by utilising methods like maximum pooling or average pooling, which take the maximum or average values within localised regions. CNN uses pooling to generalise the characteristics that the convolutional filters have retrieved, allowing the network to recognise patterns regardless of where they are in the image.

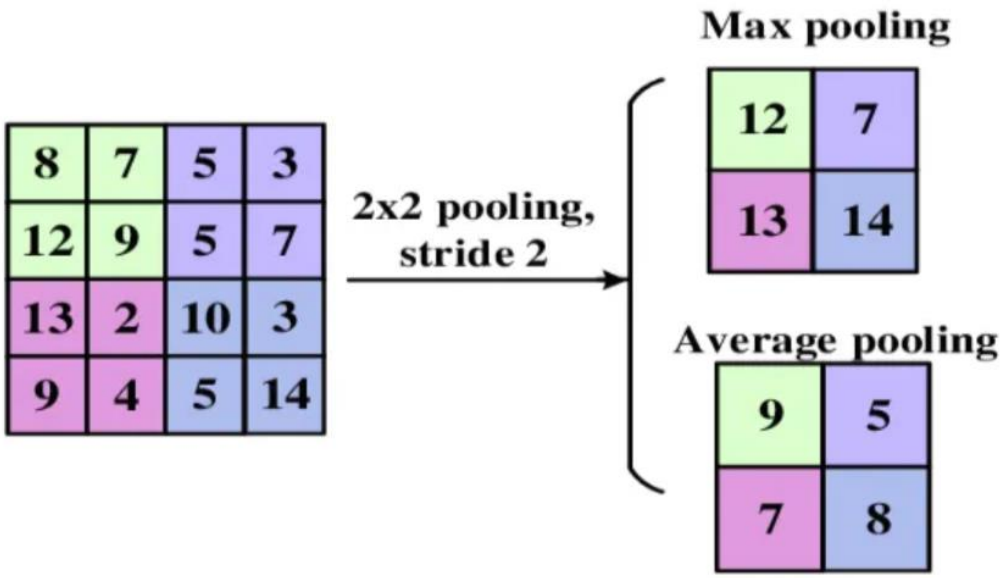

Fig2.5 Pooling

Activation Function

Activation functions play a crucial role in determining the activity level and contribution of neurons in a neural network. They help in making predictions by introducing non-linearity to the network. Commonly used activation functions such as ReLU, Softmax, tanH, and Sigmoid serve specific purposes and enable the network to learn complex patterns and correlations in the data.

CHAPTER 3

TRAINING MODELS USED

1. **VGG 16:** VGG is famous CNN architecture. It has the convolution layer with filter size 3x3 with stride size of 1 and always have the same padding and maxpooling layer with filter size of 2x2 with stride size of 2. This configuration of convolution and max pool layers is constant throughout the design. The output is handled by a softmax and two FC (fully connected layers). The 16 in VGG16 stands for the number of weighted layers, which are 16, in the model[5].

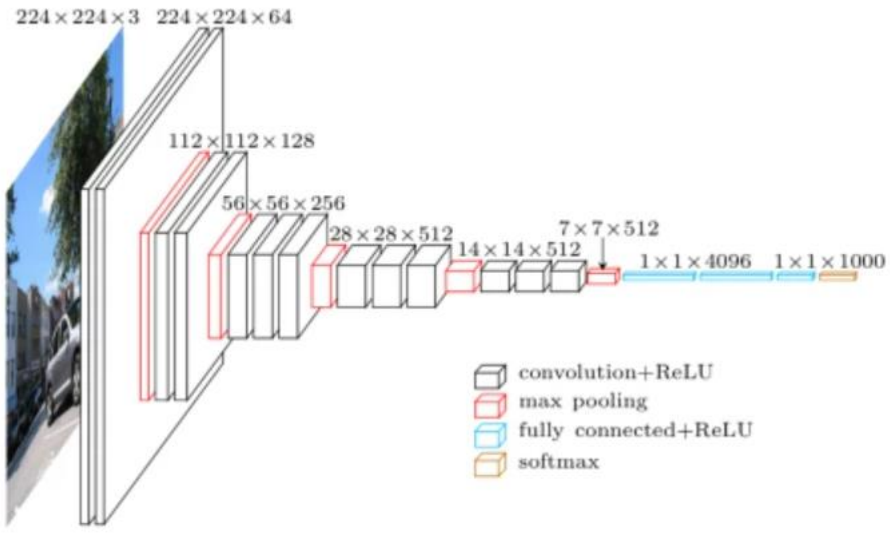

Fig 3.1 Architecture of VGG16

2. **XCEPTION**

   Deep CNN architecture called Xception uses depthwise separable convolutions. One way to think of a depthwise separable convolution is as an Inception model with a maximum number of towers.

3. RESNET50

Convolutional Neural Networks (CNNs) are widely used in visual perception tasks, and one of the popular models is ResNet-50. ResNet-50 is a pre-trained deep learning model specifically designed for image recognition. It was trained on a large dataset called ImageNet, which consists of one million images belonging to 1000 different categories. With 50 layers, the ResNet-50 model is organized into five stages, each containing a residual block. These residual blocks are made up of three layers, involving 1x1 and 3x3 convolutions. By leveraging this pre-trained model, researchers and practitioners can benefit from its ability to accurately identify and classify photos

Each layer in a network with leftover blocks feeds into the layer below it.and then straight into the layers that are around 2-3 hops away. Identity links are the names given to these relationships.

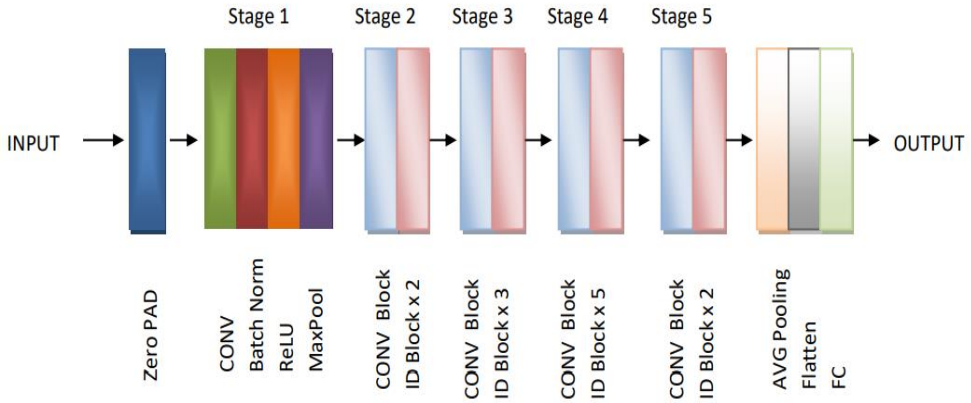

Fig3.2 Residual Learning Block and Architecture of ResNet-50

4. INCEPTION

The overfitting of the data was caused by many deep layers of convolutions. The inception model uses parallel layers or numerous filters of various widths on the same level to make the model wider rather than deeper in order to prevent overfitting**.** It has 42 layers in the InceptionV3 model.

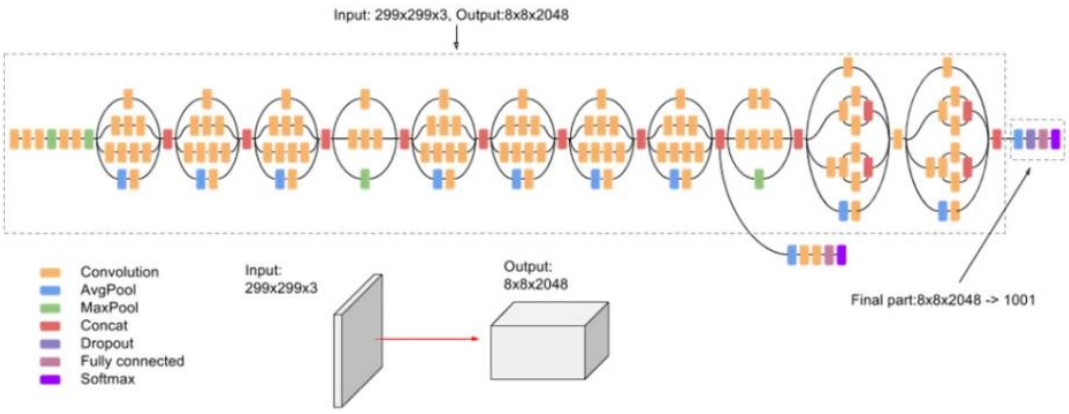

Fig 3.3 Architecture of Inception

5. INCEPTIONRESNET

A CNN model called InceptionResNet-v2 was created by Google researchers. This model aimed to simplify InceptionV3 and investigate the viability of using residual networks to the Inception model**.**

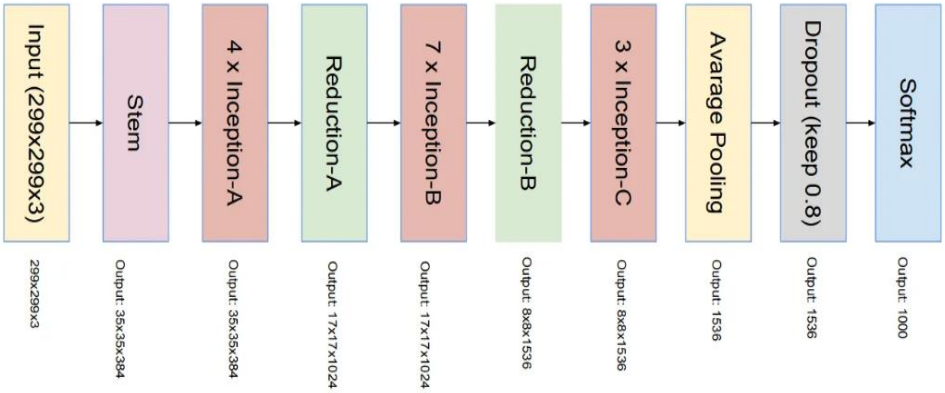

Fig 3.4 Architecture of InceptionResnet

6. VGG19

A popular classifier is the VGG19 model, a well-known deep learning architecture that can be found in the TensorFlow framework. In this work, a new head created especially for a distinct classification task was built on top of VGG19.

One pooling layer and two sets of convolutional filter layers are replicated twice in the deep neural network VGG19. In addition, one pooling layer and four sets of convolutional filter layers are applied three times each. A softmax output layer and two completely linked layers are also included in the architecture.

The final levels of the network are taken out and new layers are added in order to modify the pre-trained VGG19 model to the particular classification problem. A flatten layer is placed after a 0.2 dropout layer in the new layer design. A further 0.2 dropout layer is added after two fully linked layers with dimensions of 128 and 64, respectively, are connected.

Three class heads are included in the final layer of the new design, which is employed for the specific classification task. The model is trained and optimised using the categorical cross-entropy loss function for this purpose.

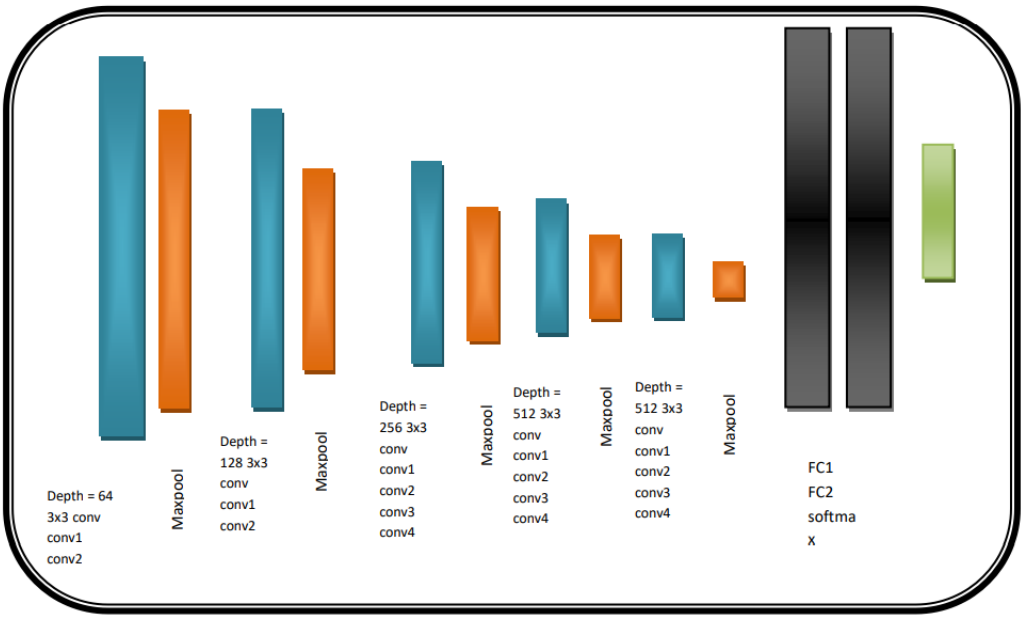

Fig 3.5 Architecture of VGG19

Chapter 4

METHODOLOGY

4.1 Dataset

- Training data The models are trained using the training dataset, and their parameters are optimised. It has several facial photos and captions for various personality traits. The models use this dataset to discover patterns and correlations between face characteristics and personality traits throughout the training phase. The models seek to reduce the mean absolute error loss function and enhance their prediction accuracy by iteratively changing their parameters using optimization techniques like Adam.
- The models are adjusted, and the top performers are chosen, using the validation dataset. It offers a diverse collection of facial photos with real-world descriptions of personality traits that were not visible during training. After each training period, the models are assessed on this dataset, and performance measures like mean absolute error or coefficient of determination are calculated. This assists in assessing the model's capacity for generalisation and preventing overfitting to the training set of data. Researchers may decide on model design, hyperparameter tweaking, and regularisation methods with confidence thanks to the validation dataset.
- The test dataset is used to evaluate the trained models overall performance. It is made up of face photos the models haven't seen during training or validation. On this dataset, the models are used, and the predicted personality characteristics are compared to the actual labels. The performance of the models and their capacity to generalise to new data are assessed objectively using the test dataset. It aids in calculating the predicted accuracy of the models and determining their applicability for practical applications.

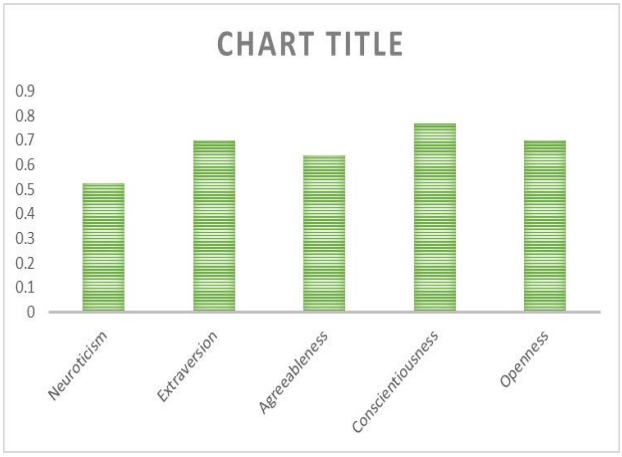

Fig.4.1 **Training dataset**

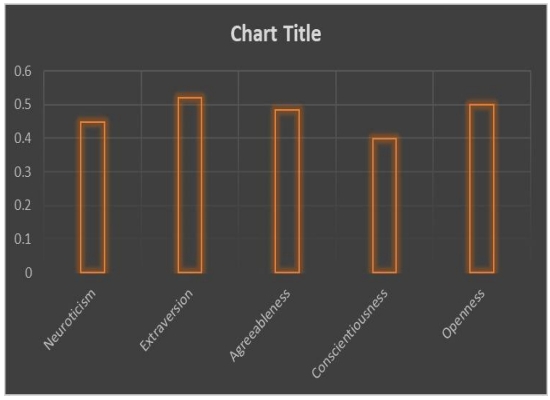

Fig.4 .2 **Validation dataset**

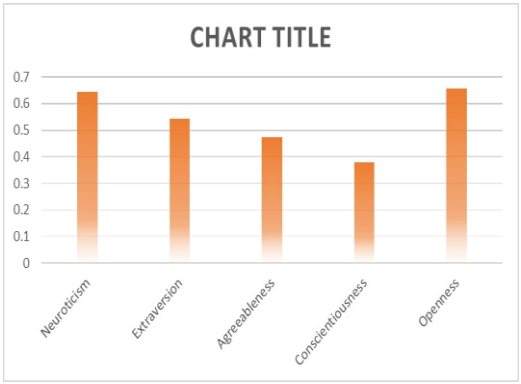

Fig.4.3 **Test dataset**

4.2 Architecture used:

1. **VGGish CNN**: VGG is a well-known Convolutional Neural Network (CNN) architecture widely used for image classification. Its key characteristic is the consistent arrangement of convolutional layers with a filter size of 3x3, a stride size of 1, and always maintaining the same padding. The architecture also includes max pooling layers with a filter size of 2x2 and a stride size of 2, which contribute to downsampling the feature maps. This pattern of convolution and max pooling layers is repeated throughout the network. VGG typically consists of two fully

connected layers (FC) and uses softmax activation for the final output. The name "VGG16" signifies the presence of 16 weighted layers in the network.

4.3 Proposed Methodology:

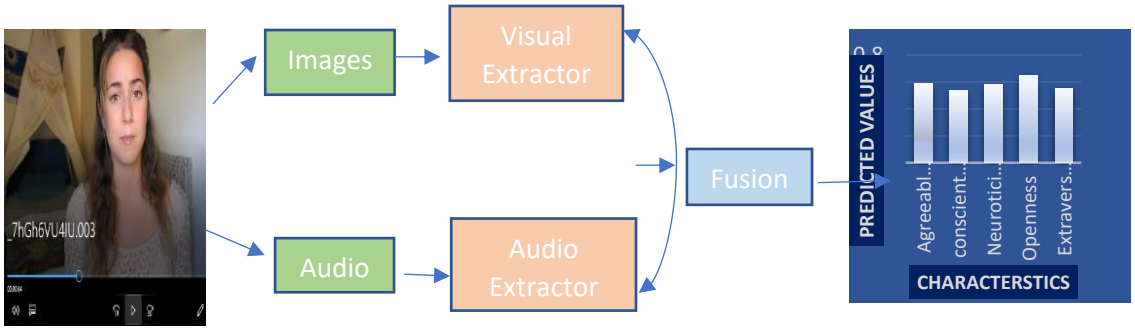

Fig.4.3 Proposed framework

The method involves extracting features from two distinct modalities, particularly video and audio. The First feature is ambient feature includes person and surrounding, the second feature is based on audio information extracted from the video, we then train them separately in a unimodal.

4.4 Visual Modality:

Facial features are extracted, it has three subparts:

- Preprocessing
- and feature extraction using CNN
- model architecture

The merging of images features to forecast the big five personality traits is the final phase.

4.4.1 PREPROCESSING:

Because of the noise in the data, it is sometimes not ideal for learning patterns in its raw form. Pre-processing of data is essential for extracting and learning relevant aspects. CNN architecture was used to extract relevant feature from video. The most well-known and widely used CNNs, such as VGGs and ResNet , take a single image

as input. The initial stage in pre-processing is to extract images from video. The **FFmpeg** is used to accomplish this. Because these images are high resolution (720p or 1080p), training is impossible; therefore, all images are scaled to 248x140 pixels. The RGB colour frame is used to extract the images.

2.4.2 Extraction of ambient features

To handle the computational and memory limitations posed by the dataset, certain optimizations were implemented. Given the average video duration of 15 seconds and the high frame rate of 30 frames per second, training the model on every frame would require significant computational resources. To address this, only six equally spaced images were extracted from each video, significantly reducing the amount of data while still providing representative frames. This downsampling strategy reduced the RAM requirement to approximately 20 GB, making it feasible for training. The extracted frames were further processed by downsampling to about 100 frames per video, ensuring a manageable dataset size. During training, a random crop of 128x128 pixels was used, while the validation and test sets employed a central crop of the same size. Each frame was labeled with the corresponding video's Big Five trait values, enabling the training of convolutional neural networks (CNNs) to develop deep regressors for personality analysis based on the extracted frames.

4.4.2 Create the visual subnetwork

VGG like architecture is used which has 2D convolutional layer and Max Pooling layer.

Visual model takes input size where “6” is the number of frames, “128 ,128” are the spatial dimensions. In Time Distributed layer, Conv2D blocks that are formed will be trained for our intended detection, allowing our image to be processed to identify anything that changing from frame to frame, rather than a basic object recognition. LSTM stands for "Long Short Time Memory," and it's a layer that can take many chronological inputs and predict what's important. As the images are in sequential order LSTM which is able to use memory is the perfect model to detect relation between images to images in a given time.

4.5 Audio Modality

Audio features are extracted, it has three subparts:

- Preprocessing
- and feature extraction using CNN
- model architecture

4.5.1 Audio representation

Using FFmpeg the audio is extracted from the video. We initially remove the audio characteristics from the source videos before learning a linear regressor for the audio modality. The audio from the original videos is extracted using the free and open-source FFmpeg software. In terms of FFmpeg's options, we select two channels for the audio outputs in WAV format, 44,100Hz for the sampling frequency.

4.5.2 Pre-processing the data

- Extract the audio from the video using ffmpeg: (Fast Forward Moving Picture Experts Group): The powerful multimedia framework FFMPEG is capable of performing a variety of audio and video processing tasks. It is a library that handles video, audio and other media files.
- After removing the audio from the movie and saving it as an AAC file, we can compute the MFCCs in Python using the librosa module. Popular Python library for audio and music analysis is called Librosa. It offers a variety of audio processing options, including MFCC extraction.
- iterated through all of the audio files in your dataset and figure out how many frames there are in each file before you can determine the average number of frames in the dataset. The mean of these values can then be determined.
- In order to standardise the MFCCs, the data must be transformed to have zero mean and unit variance. The characteristics should be normalised and made similar across samples using this procedure.
- Applying pre-padding to the standardised MFCCs will uniformly lengthen the audio samples. Pre-padding entails lengthening each sample to the same length by commencing it with zeros.

The final audio representation for each sample is a tensor of the form (N, M, 1), where N denotes the number of coefficients (for instance, 24), M is the number of audio frames (corresponding to the maximum length after pre-padding), and 1 denotes the single channel (as we are working with mono audio).

4.5.3 Create the audio subnetwork

- The CNN and Alexnet which consists of convolutional layers followed by pooling layers, processes the audio inputs. These layers are in charge of recording the audio's pitch, frequency, and temporal patterns, among other characteristics. The model obtains high-level representations that encode significant information about the audio input through this method.
- Following feature extraction, a fully connected layer and sigmoid activation function are used to the audio features to produce the final output prediction. The system uses fusion techniques like concatenation or attention processes to merge these audio information with features from other modalities, such visual modality.
- The multi-modal personality prediction method may accurately capture the distinctive features of a person's personality displayed through their voice, tone, and speech patterns by adding the audio modality. This all-encompassing strategy improves the system's capacity for prediction and offers a better comprehension of a person's personality based on several modalities.

4.6 COMBINING MODALITIES

Latent features from two different modalities are integrated, and the resulting feature for the sample is given to a fully connected layer. As a result, the created model is a network from end to end. The sigmoid layer receives the composite feature at the end for trait prediction.

4.7 TRAINING

The network was trained using the Adam optimizer by iteratively minimizing the mean absolute error loss function for 50 epochs. This process involved continuously updating the model's parameters to improve the predictions and align them with the desired goal traits. The network was optimised using the Adam optimizer while the multimodal

personality prediction system was in the training phase. Iteratively reducing the mean absolute error loss function was the goal of the optimization procedure. This involves modifying the model's parameters across a number of epochs to enhance prediction precision and match the goal attributes. The system successfully changed the weights and biases of the network using the Adam optimizer, enabling the convergence to the best prediction performance.

```
 dense_12 (Dense)               (None, 128)          524416      ['dropout_1[0][0]']

 lstm_9 (LSTM)                  (None, 64)           409856      ['time_distributed_9[0][0]']

 concatenate_11 (Concatenate)   (None, 192)          0           ['dense_12[0][0]',
                                                                  'lstm_9[0][0]']

 dense_23 (Dense)               (None, 256)          49408       ['concatenate_11[0][0]']

 dense_24 (Dense)               (None, 5)            1285        ['dense_23[0][0]']

==================================================================================================
Total params: 365,429,701
Trainable params: 365,428,485
Non-trainable params: 1,216
```

```
 conv2d_10 (Conv2D)             (None, 5, 329, 384)  885120      ['batch_normalization_9[0][0]']

 conv2d_11 (Conv2D)             (None, 5, 329, 384)  1327488     ['conv2d_10[0][0]']

 conv2d_12 (Conv2D)             (None, 5, 329, 256)  884992      ['conv2d_11[0][0]']

 max_pooling2d_10 (MaxPooling2D  (None, 2, 164, 256)  0          ['conv2d_12[0][0]']
 )

 batch_normalization_10 (BatchN  (None, 2, 164, 256)  1024       ['max_pooling2d_10[0][0]']
 ormalization)

 flatten_2 (Flatten)            (None, 83968)        0           ['batch_normalization_10[0][0]']

 dense_10 (Dense)               (None, 4096)         343937024   ['flatten_2[0][0]']

 dropout (Dropout)              (None, 4096)         0           ['dense_10[0][0]']

 dense_11 (Dense)               (None, 4096)         16781312    ['dropout[0][0]']

 input_38 (InputLayer)          [(None, 6, 128, 128  0           []
                                , 3)]

 dropout_1 (Dropout)            (None, 4096)         0           ['dense_11[0][0]']

 time_distributed_9 (TimeDistri  (None, 6, 1536)     0           ['input_38[0][0]']
 buted)

 dense_12 (Dense)               (None, 128)          524416      ['dropout_1[0][0]']
```

CHAPTER 5

Model Evaluation: For evaluating predictions, we employ the following metrics based on confusion matrices.

3.1 Coefficient of Determination (R2)

The coefficient of determination, also known as R-squared (R2), is a statistical measure that quantifies the proportion of the dependent variable's variability that can be explained by the independent variable(s) in a regression model. It provides an indication of the goodness of fit of the regression model to the observed data, showing how well the model predicts the variation in the dependent variable.

1. Epoch= 10

Coefficient Determination (R^2)

```
]: from sklearn.metrics import r2_score

]: pers = ['Neuroticism','Extraversion','Agreeableness','Conscientiousness','Openness']
   r2s = [r2_score(test_input[2][:,i,:], normalized[:,i,:]) for i in range(5)]
   for pers, r2 in zip(pers,r2s):
       print(pers + ': ' + str(r2))

Neuroticism: -16.023398513951218
Extraversion: -60.56663713749815
Agreeableness: -117.40300492885557
Conscientiousness: -59.001268217492
Openness: -569.8051676622218
```

2. Epoch= 20

Coefficient Determination (R^2)

```
: from sklearn.metrics import r2_score

: pers = ['Neuroticism','Extraversion','Agreeableness','Conscientiousness','Openness']
  r2s = [r2_score(test_input[2][:,i,:], normalized[:,i,:]) for i in range(5)]
  for pers, r2 in zip(pers,r2s):
      print(pers + ': ' + str(r2))

Neuroticism: -32.164004149804526
Extraversion: -16.18052118607601
Agreeableness: -32.898527293580585
Conscientiousness: -6.4071271479426235
Openness: -18.015577469193932
```

3. **Epoch 20 when Convolution3D is used in Visual Subnetwork**

Coefficient Determination (R^2)

```
In [54]: from sklearn.metrics import r2_score

In [55]: pers = ['Neuroticism','Extraversion','Agreeableness','Conscientiousness','Openness']
         r2s = [r2_score(test_input[2][:,i,:], normalized[:,i,:]) for i in range(5)]
         for pers, r2 in zip(pers,r2s):
             print(pers + ': ' + str(r2))

Neuroticism: -2.719354887580563
Extraversion: -4.2321965255594645
Agreeableness: -5.774594334598174
Conscientiousness: -2.8780059781499814
Openness: -37.8458287485802
```

4. **Epoch=50**

Coefficient Determination (R^2)

```
In [117]: from sklearn.metrics import r2_score

In [118]: pers = ['Neuroticism','Extraversion','Agreeableness','Conscientiousness','Openness']
          r2s = [r2_score(test_input[2][:,i,:], normalized[:,i,:]) for i in range(5)]
          for pers, r2 in zip(pers,r2s):
              print(pers + ': ' + str(r2))

Neuroticism: -0.18698877538453806
Extraversion: -0.47895672198140593
Agreeableness: -0.5604523589792347
Conscientiousness: -0.2061734690310062
Openness: -0.3581858075115578
```

3.2 MAE: Mean Absolute Error is used to evaluate predict. The average magnitude of errors between anticipated and actual values is measured using this metric, which is frequently employed in regression analysis.

The following equation can be used to determine MAE:

$MAE = (1/n) * \Sigma|y_i - \hat{y}_i|$

MAE determines the average absolute difference between predicted and actual values, which reflects how close, on average, the predictions are to the actual values. The MAE number is always non-negative, and a lower MAE indicates better quality.

Accuracy can be calculated as: -

Accuracy = 1-MAE

Table 5.1: Accuracy Evaluation for Various Audio-Visual Models (1-MAE)

Modality	**Models**	**Accuracy (1-MAE)**
Audio-visual	Alexnet+VGG19	0.911
Audio-visual	Alexnet+VGG16	0.889
Audio-visual	CNN+VGG16	0.887
Audio-visual	Alexnet+Xception	0.884
Audio-visual	Alexnet+InceptionResnetV2	0.882
Audio-visual	Alexnet+InceptionV3	0.88
Audio-visual	Alexnet+Convnext	0.879
Audio-visual	CNN+Resnet50	0.874

Our approach shows slightly higher accuracy than the base model Berkay Aydin[3]

This paper shows the accuracy using and employed a random decision forest regression method for predicting the Big Five personality traits.

Table 5.2: Comparison With Baseline Model

Model	Accuracy
Random Decision Tree	0.899
Alexnet+VGG19	0.91

CONCLUSION

In this paper, we provide a framework for automatically predicting personality traits from brief videos. For the Chalearn First Impressions Challenge, the approach was created. We have created a method that makes use of acoustic and visual elements. A comprehensive examination of all the features reveals that almost all of the retrieved features behave similarly, achieving accuracy levels between 87% and 91%.

Through our studies, we assessed the accuracy of several audio-visual architectures in predicting the Big Five personality characteristics. According to the findings, the audio-visual architecture that included AlexNet and VGG19 had the best accuracy (0.91). This shows that using both visual and aural modalities for personality prediction is beneficial.

In summary, our work shows the value of pre-processing data to extract pertinent aspects for successfully learning patterns. We were able to create a strong framework for predicting the Big Five personality traits by using CNN architectures and relying on known models like VGG19. The model's prediction ability was further improved by the addition of audio and visual modalities. This study is a contribution to the field of personality analysis and sheds light on how deep learning models may be used to comprehend and forecast human behaviour using multimodal data**.**

REFERENCES

[1] Suman, Chanchal, Sriparna Saha, Aditya Gupta, Saurabh Kumar Pandey, and Pushpak Bhattacharyya. "A multi-modal personality prediction system." *Knowledge- Based Systems* 236 (2022): 107715.

[2] Singh, Brijesh Kumar, Mansi Katiyar, Shefali Gupta, and Nikam Gitanjali Ganpatrao. "A Survey on: Personality Prediction from Multimedia through Machine Learning." In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 1687-1694. IEEE, 2021.

[3] Aydin, Berkay, Ahmet Alp Kindiroglu, Oya Aran, and Lale Akarun. "Automatic personality prediction from audiovisual data using random forest regression." In *2016 23rd international conference on pattern recognition (ICPR)*, pp. 37-42. IEEE, 2016.

[4] Wei, Xiu-Shen, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. "Deep bimodal regression of apparent personality traits from short video sequences." *IEEE Transactions on Affective Computing* 9, no. 3 (2017): 303-315.

[5] Rohit, G. V., K. Rakesh Bharadwaj, R. Hemanth, Bariti Pruthvi, and MV Manoj Kumar. "Machine intelligence based personality prediction using social profile data." In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pp. 1003-1008. IEEE, 2020.

[6] Chen, Yu, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. "Adversarial posenet: A structure-aware convolutional network for human pose estimation." In *Proceedings of the IEEE international conference on computer vision*, pp. 1212-1221. 2017.

[7] W. M. K. S., and T. G. I. Fernando. "Performance Analysis of State-of-the-Art Deep Learning Models in the Visual-Based Apparent Personality Detection." *Vidyodaya Journal of Science* 25, no. 02 (2022).

[8] Sun, Xiao, Jie Huang, Shixin Zheng, Xuanheng Rao, and Meng Wang. "Personality Assessment Based on Multimodal Attention Network Learning With

Category-Based Mean Square Error." *IEEE Transactions on Image Processing* 31 (2022): 2162-2174.

[9] Mandıra, Burak, Dersu Giritlioglu, Selim Fırat Yılmaz, Can Ufuk Ertenli, Berhan Faruk Akgür, Merve Kınıklıoglu, Aslı Gül Kurt et al. "Spatiotemporal and Multimodal Analysis of Personality Traits." In *15th International Summer Workshop on Multimodal Interfaces*, p. 32. 2019

[10] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[11] Tadesse, Michael M., Hongfei Lin, Bo Xu, and Liang Yang. "Personality predictions based on user behavior on the facebook social media platform." *IEEE Access* 6 (2018): 61959-61969.

[12] Mehta, Yash, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. "Recent trends in deep learning based personality detection." *Artificial Intelligence Review* 53 (2020): 2313-2339.

[13] Güçlütürk, Yağmur, Umut Güçlü, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel AJ Van Gerven, and Rob Van Lier. "Multimodal first impression analysis with deep residual networks." *IEEE Transactions on Affective Computing* 9, no. 3 (2017): 316-329.

[14] Thilakaratne, Menasha, Ruvan Weerasinghe, and Sujan Perera. "Knowledge-driven approach to predict personality traits by leveraging social media data." In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 288-295. IEEE, 2016.

[15] Escalante, Hugo Jair, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Gucluturk, Umut Guclu, Xavier Baró et al. "Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos." *arXiv preprint arXiv:1802.00745* (2018).

[16] Gorbova, Jelena, Iiris Lusi, Andre Litvin, and Gholamreza Anbarjafari. "Automated screening of job candidate based on multimodal video processing." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 29-35. 2017.

turnitin

Similarity Report ID: oid:27535:36496334

PAPER NAME

anisha.pdf

WORD COUNT

6879 Words

CHARACTER COUNT

38991 Characters

PAGE COUNT

36 Pages

FILE SIZE

1.1MB

SUBMISSION DATE

May 30, 2023 12:51 PM GMT+5:30

REPORT DATE

May 30, 2023 12:52 PM GMT+5:30

7% Overall Similarity

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

- 5% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 4% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less then 10 words)