CROSS-AGE EFFECT IN ARTIFICIAL NEURAL NETWORKS: A STUDY ON FACIAL AGE RECOGNITION BIAS IN ARTIFICIAL NEURAL NETWORKS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
MIDDLE EAST TECHNICAL UNIVERSITY
BY

NECATİ ÇAĞATAY GÜRSOY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COGNITIVE SCIENCES

JUNE 2019

CROSS-AGE EFFECT IN ARTIFICIAL NEURAL NETWORKS: A STUDY ON FACIAL AGE RECOGNITION BIAS IN ARTIFICIAL NEURAL NETWORKS

Submitted by Necati Çağatay Gürsoy in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Sciences Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science** _____

Asst. Prof. Dr. Murat Perit Çakır
Supervisor, **Cognitive Science Dept., METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Cengiz Acartürk
Cognitive Science Dept., METU _____

Asst. Prof. Dr. Murat Perit Çakır
Cognitive Science Dept., METU _____

Asst. Prof. Dr. Özkan Kılıç
Computer Engineering Dept., Ankara Yıldırım Beyazıt University _____

Date: *_27th of June, 2019*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Necati Çağatay Gürsoy

Signature : _____

ABSTRACT

CROSS-AGE EFFECT IN ARTIFICIAL NEURAL NETWORKS: A STUDY ON FACIAL AGE RECOGNITION BIAS IN ARTIFICIAL NEURAL NETWORKS

Gürsoy, Necati Çağatay

MSc., Department of Cognitive Sciences

Supervisor: Assist. Prof. Dr. Perit Murat Çakır

June 2019, 80 pages

Simply put cross-age effect or own-age bias is the phenomenon that when guessing the age of individuals from their faces, it is claimed that the age of the individual who is guessing the facial age influences the guessing process, and the effect is towards to the age of the one who is guessing. Previous studies on the phenomenon acknowledges that such bias exists, and especially in forensics this bias can cause drastic faults. In our two-part study, firstly the phenomenon is investigated in humans to observe if such an effect exists, then it is tested if the error could be reduced via utilizing convolutional neural networks on face images and classify the face images with respect to their ages. For the human based experiment, participants rated facial images and for the neural network based experiment a convolutional neural network model was constructed and tested with same images. In human experiments, a strong correlation between participants' guesses and the real ages of the face images was observed. Correlation analysis yield that a positive relationship between error (guessed age – actual age) and objective distance (participant's age – actual age) exists, whereas almost zero correlation exists between error and perceived distance (participant's age – guessed age). Neural network experiment indicated that the neural network's age rating performance exceeded human performance. Moreover; it was claimed that with necessary pre-processing to the input data, cross-age effect can be deduced with neural networks and human error can be reduced significantly.

Keywords: Own-age bias, cross-age effect, artificial neural networks, facial age recognition, facial age memory

ÖZ

YAPAY SİNİR AĞLARINDA YAŞLAR ARASI ETKİ: YAPAY SİNİR AĞLARINDA YÜZSEL YAŞ TANIMLAMA SAPMASINA DAİR BİR ARAŞTIRMA

Gürsoy, Necati Çağatay

Yüksek Lisans, Bilişsel Bilimler Bölümü

Tez Yöneticisi: Dr. Öğr. Üyesi Perit Murat Çakır

Haziran 2019, 80 sayfa

Yaşlar arası etki veya kendi-yaş sapması kişilerin yüzlerine bakılarak yaş tahmini yapılırken, tahmin yapan kişinin kendi yaşının yaptığı tahmini etkilediğini ve etki yönünün kişinin kendi yaşına doğru olduğunu öne süren bir görüngüdür. Bu görüngü üzerindeki önceki araştırmalarda bahsedilen etkinin var olduğu ve özellikle Adli Tıp'ta büyük hatalara yol açtığı belirtilmiştir. İki parçalı araştırmamızın ilk kısmında bu görüngü insanlar üzerinde iddia edilen etkinin var olup olmadığı yönünde sınanmıştır. Daha sonra bu etkinin yol açtığı hataların en aza indirilebilmesi için evrişimli sinir ağları kullanılarak yüz görüntüleri yaşlara göre sınıflandırılmıştır. İnsanlarla yapılan deneyde, katılımcılar yüz görüntülerinin yaşlarını tahmin etmiştir, yapay sinir ağları üzerinden yapılan deneyde ise bir evrişimli sinir ağları modeli gerçeklenerek aynı yüz görüntülerinin yaş tahminleri alınmıştır. İnsanlarla yapılan deneyde, katılımcıların yaş tahminleri ile yüz görüntülerinin asıl yaşları arasında güçlü bir doğrusal ilişki tespit edilmiştir. Korelasyon analizi yapılarak esas hata (tahmin edilen yaş – gerçek yaş) ve nesnel uzaklık (katılımcının yaşı – gerçek yaş) arasında pozitif bir ilişki bulunduğu tespit edilmiştir. Yapay sinir ağları deneyinde ise modelin yaş tahmin isabet performansının insan katılımcıların performansının üzerine çıktığı tespit edilmiştir. Ayrıca; gerekli girdi verisi ön-hazırlığı ile yaşlar arası etkisinin yapay sinir ağları yardımıyla azaltılabileceği ve insan hatasının önemli miktarda azaltılabileceği iddia edilmiştir.

Anahtar Sözcükler: Kendi-yaş sapması, yaşlar arası etki, yapay sinir ağları, yüzsel yaş tanıma, yüzsel yaş hafızası

ACKNOWLEDGMENTS

Firstly, I would like to thank my thesis advisor Dr. Murat Perit Çakır for he has given me the opportunity to work in this ever-young subject of Cognitive Science. Dr. Çakır had always had an open door for me, whenever I had run into a trouble during our advisor-pupil relationship he had cleared the fog of thought in my brain and directed me the way that I should continue. Moreover, when I came up with the odd research question that had evolved into this thesis, he did not turn me back. Instead he pointed me the essential points that would explain the question.

Additionally, I am grateful to Dr. Cengiz Acartürk for inspiring and supporting me on the baby steps of my journey to become a scientist. For Dr. Acartürk have shown me the way to critical and objective thinking.

Furthermore, Dr. Tolga Esat Özkurt have re-shaped my outlook on Neuroscience in a more joyous and deeper way. I would like to thank him for being a guide for my path to acquiring more and more cognition to realize that the ultimate knowledge is actually "knowing nothing at all!".

Also, I wanted to show my gratitude to all my instructors and fellow pupils in METU Cognitive Science department for they have presented an open way of arguing and contributing to science.

Finally; I would like to thank my parents, Sibel and Hüseyin and my brother Çağkan for their invaluable support in the first step of my aspiration to become an "*academic*". Moreover, I would like to thank Toygar and Akarsu whom both contributed with my journey to the land of "knowledge is knowing nothing!" with their endless discussions.

Lastly, I would never be thankful enough for the eternal support and love of my dear Özge; I cannot imagine where would I be without her counselling. May both our journeys in science will be always in our favour.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
ANOVA	Analysis of Variances
AUC	Area Under the Curve
CAE	Cross-Age Effect
CIM	Categorization-Individuation Model
CM	Confusion Matrix
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CRE	Cross-Race Effect
Deep-Net	Deep Learning Network
DL	Deep Learning
EEG	Electroencephalography
ERP	Event Related Potential
FG-NET-AD	FG-NET Aging Database
FRGC	Face Recognition Grand Challenge
GPU	Graphical Processing Unit
HTML	Hyper Text Markup Language
METU	Middle East Technical University
ML	Machine Learning
MySQL	"My" Structured Query Language
OAB	Own-Age Bias
OCR	Optical Character Recognition
ORB	Own-Race Bias
PAML	Park Aging Mind Lab at University of Texas Dallas
PAML-AD	Park Aging Mind Lab at University of Texas Dallas Aging Database
PHP	Personal Home Page OR PHP: Hypertext Preprocessor
ROC	Receiver Operating Characteristics
SPSS	IBM Statistical Package for the Social Sciences
SVM	Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Face Recognition

As humans, we live in a domain filled with various unique visual stimuli classes, such as chairs, tables, pianos, guitars, cars, apartment buildings, trousers, shoes, hats and such examples may go on and on. Processing all those stimuli classes is handled by featural mechanisms, which means that, we identify all objects by bonding all its atomic features together in order to create a combined visual representation of a larger entity; i.e. for a piano, atomic features would be keys, hammers, strings, pedals etc (Biederman, 1987). This holistic approach to visual perception can be applied to face recognition as well; since any face is a unique combinatory entity of smaller entities, that are eyes, nose, mouth, ears, hair and any facial feature that one can visualize. This holistic processing assists humans in perceiving faces as a "Gestalt-like" structure where the whole face is not simply the combination of its atomic parts. Rather this is a complex structure which includes sensical and semantical connections in extent to atomic parts (Rivolta, 2014).

As reported in Rivolta (2014), two prominent hypotheses about facial recognition processes are widely accepted; namely the *domain-specific hypothesis* and the *expertise hypothesis*. Researchers that support the domain-specific hypothesis have designed experiments that would enlighten the atomic parts and their interconnections in domain-specific hypothesis. In such experiments, participants are typically instructed to learn different identities, i.e. humans, and then they were asked to memorize that main identity's features, i.e. eyes. Lastly, participants are required to distinguish main identities' features from each other. For instance, the experiment may require a participant to indicate if the identity's chosen feature is in the right or left image that was shown. The results of such experiments implied that the identification performance was better when the features of the main identity were isolated from the main entity (i.e. if the eyes of the human 1 was replaced with human 2's eyes). This effect is called as the *part-whole effect*, which diminishes in horizontally inverted faces and non-face objects. Hence, these finding imply that such effects can only play an important role in non-inverted faces as the facial contours signal the correct recognition of facial features i.e. eyes (Rivolta, 2014; Tanaka & Farah, 1993).

Figure 1: The *part-whole effect* demonstration from Tanaka and Farah's experiments. Top images were the ones where the chosen feature is facilitated in the face configuration whereas bottom images are features in isolated images. Illustration taken from Rivolta (2014) (Rivolta, 2014, p.22; Tanaka & Farah, 1993).

Although there are various experiments that support the domain-specific hypothesis, the expertise hypothesis has arisen as a critique of the reductionist approach favoured by the domain-specific view where face processing is viewed from a holistic perspective. The hypothesis simply claims that during development of holistic mechanisms for face recognition, expertise on face classification plays a significant role. Moreover, it claims that the depth of processing creates the difference on recognizing between faces and non-face objects, which means when a non-face object is seen a less detailed identification process is handled as compared to observing a face. Early studies on this hypothesis had somewhat satisfying results, but recently there were studies which seriously disproves the hypothesis. One such experiment was reported by Robbins and McKone (2007), which included two groups; dog-experts (breeders, trainers or dog judges who have an average of 23 years' experience with dogs) who are expert in classifying dogs into groups and people who are not dog-experts. Both of the experimental groups were supposed to be experts in face perception in theory. In the experiment, both groups were tasked to memorize upright dog and face images. Then, they were subjected to memorized stimuli with distractive stimuli, where they were required to choose the actual stimuli. This procedure was conducted again, but this time with horizontally rotated stimuli. The results have shown that participants who are not dog-experts have been affected by this rotation effect, i.e. their face image recognition performance significantly dropped with inverted stimuli when compared to dog images. Moreover, even though they were considered as *experts*, dog-experts have shown larger face inversion effect when compared to inversed dog images. This finding alone discourages the expertise hypothesis and supports the domain-specific hypothesis, as it is clear that each face had uniqueness to some extent, and being an expert on recognition did not aided individuals in recognizing warped stimuli (Rivolta, 2014; Robbins & McKone, 2007).

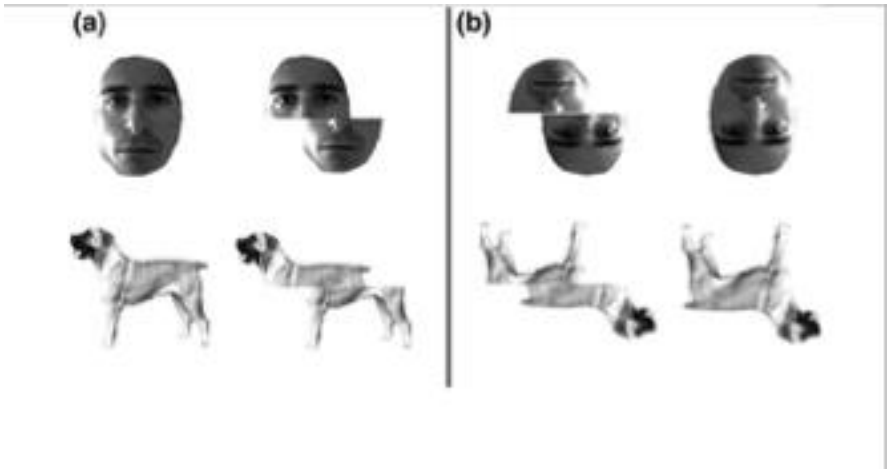

Figure 2: Representational stimuli that was used in the Robbins and McKone's experiment. In (a) face and dog images were displayed as aligned and misaligned conditions. In (b) same images were displayed as horizontally flipped versions. The images were created by Rivolta to represent Robbins and McKone's experiment (Rivolta, 2014, p.25; Robbins & McKone, 2007).

When human visual perception is considered, facial recognition is the expertise shared by all humans that are involved in social stimuli. This facial perception is the natural extent of humans' attention strategy, which may be simplified as having higher attention to faces or face-like objects when compared to any other objects. And this higher attention phenomenon is well demonstrated by empirical studies (Bindemann, Doherty, Burton, Schweinberger, & Langton, 2007; Hershler & Hochstein, 2005). Moreover, this higher attention is extended in the form of a well-established phenomenon that individuals are more accurate at remembering faces that are in their own group when compared to a different group (Malpass & Kravitz, 1969; Meissner & Brigham, 2001). At first, races of the individuals were the pivotal peer-groups that was thought to construct such groups. Yet in the recent years more and more peer-group descriptions were to be constructed; such as age, gender and even species (Hole & Bourne, 2010).

For our research purposes, especially the peer-group description regarding age was a pivotal point. This description can be interpreted as individuals having higher performance in recognizing faces that are already coded via individuals as being in their own age groups (Rhodes & Anastasi, 2012). This phenomenon has several logical explanations, such as individuals might be highly motivated to put attention on faces that are in their own age group during the learning phase, hence this may lead to a higher performance in recognizing in faces that are in their own age group (Hugenberg, Young, Bernstein, & Sacco, 2010).

1.2 Cross-Age Effect

The effect which individuals tend to recognize faces that they consider as in their own-group was discussed before. In short, ORB is a face recognition bias where individuals are generally better at recognizing faces that they consider as their own-race when compared to faces that they do not consider as their own-race. The own-age bias or the cross-age effect is the extension of a similar phenomenon; cross-race effect or own-

race bias which was often discussed in the light of the contact hypothesis. (Meissner & Brigham, 2001).

One of the main theoretical discussion points was that higher accuracy in face recognition may be the inevitable result of increased contact and expertise on perception of an individual's own-group faces. Several studies have reported that; with enough amount of exposure to other-group individuals, the effect may diminish. In their article where the neural aspects of other-race face recognition was investigated, Tanaka and Pierce (2009) gives two instances for this diminishing effect. In one account, on Caucasian individuals who live in a multiracial neighbourhood the CRE was reduced when compared to control participants who were in the same racial group with the faces that that they were displayed (Chiroro & Valentine, 1995). Similarly, when the CRE was tested on Caucasian individuals who watch sports events played dominantly by African American athletes. When compared to control group, the CRE was diminished significantly in contrast to the control group (MacLin, Van Sickler, MacLin, & Andrew, 2004). Yet, Tanaka and Pierce present some counter arguments as well. They report that, the interracial contact might not necessarily insure reduced CRE. For example in a related research, Caucasian individuals with extensive contact to Chinese individuals did not show improved CRE when recognizing faces from the other races (ng & Lindsay, 1994; Tanaka & Pierce, 2009).

The CAE or OAB is defined as an extension of such preferential processing of own-group faces that are relative to the faces of other-group individuals (Rhodes & Anastasi, 2012). Researchers also investigated the cognitive background of the phenomenon. Dakin and Watt (2009) used a filtering method to selectively remove all visual information but those restricted to certain orientation ranges and hence simulating what information would be passed by V1 neurons (Dakin & Watt, 2009). V1 neurons were historically called as "striate cortex" and this region is a part of all mammalian brain where many retinal fibres first arrive in the cortex. One of the key features of the neurons in this area is that each cell deals with a limited part of retina. Hence, each cell would have their own receptive field, a part of retinal cells that processes images from their own "window of the world". Such receptive fields would work as a "bar detector" and an example to this bar detector can be observed in Figure 3 (Frisby & Stone, 2010). With their experiments Dakin and Watt shown that there exists a quantitative superiority in face recognition sensitivity on horizontal facial structures when compared to facial structures with different alignments. Moreover, they imply that those horizontal structures tend to be in clusters of vertical alignments; a phenomenon that is unique for faces, which did not apply to any other natural scenery or images (Dakin & Watt, 2009).

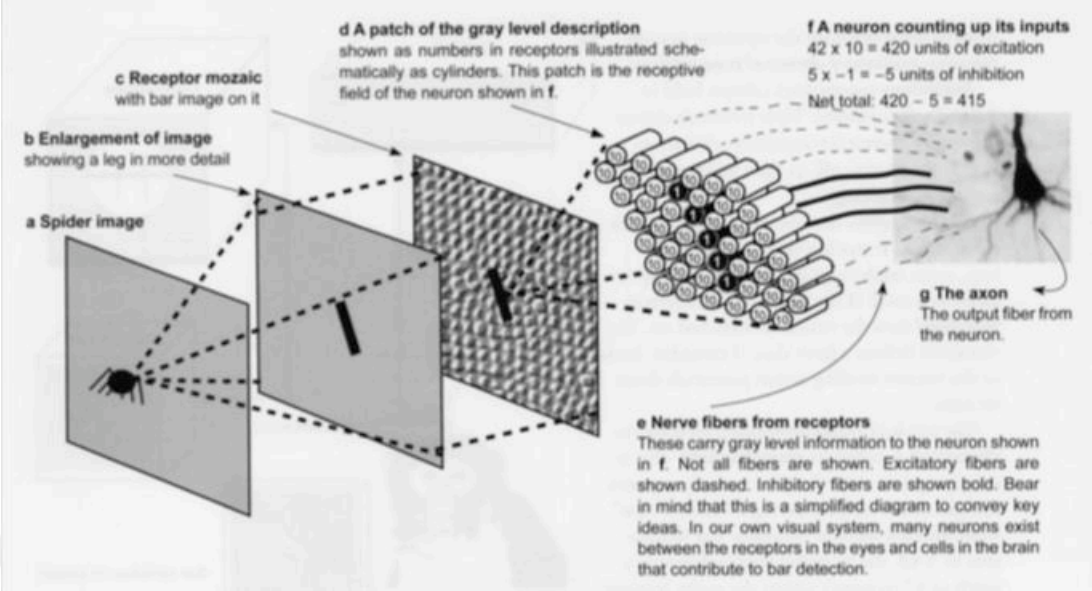

Figure 3: Bar detector design using a template implemented as a receptive field of the off-centre type. Taken from Frisby and Stone's book on biological vision (Frisby & Stone, 2010, p.57).

To investigate the electrophysiological aspect of face recognition proficiency and to possibly discover a link between CAE and neural activities, EEG studies were conducted as well. When individuals who have high and low face recognition performance was compared higher performance was associated with greater amplitudes in P100 component, which means that a positive peak approximately 100ms after stimulus onset was recorded (Turano, Marzi, & Viggiano, 2016). P100 effect is an ERP that is often linked with attention allocation and image processing (Luck, 2005). Turano et al. (2016) also reported that N170, which is an ERP component associated with face structural encoding, is critically reinforced in reaction to recognised faces in individuals with higher recognition ability when compared with individuals with lower recognition ability (Turano et al., 2016). In addition to this; it was reported that N250, an ERP linked with implicit facial recognition as well as P600 or the Late Positive Component, an ERP linked to explicit facial recognition occurred in weaker amplitudes in individuals who are suffering with developmental prosopagnosia, which is a lifelong medical condition that impairs an individual's facial recognition ability but who do not have any level of impairment in intelligence and/or memory performance with respect to healthy controls (Towler, Gosling, Duchaine, & Eimer, 2012).

Both N250 and P600 show parallelism with CAE. With individuals in young adult age group, N250 becomes greater in amplitude while reacting to repetitive observation of in-group faces when compared to out-group faces. Also, it was reported that P600 was in greater amplitude for out-group faces which potentially implies that out-group faces require higher processing when compared to in-group faces. This effect was solely reported in participants who were in young adult age groups and stated that they have little exposure to individuals that are member of other age group. Moreover, the reported effect was in line with the essential exposure basis of CAE. It is necessary to state that enhanced magnitudes of P600 were recorded in young adult participants who

correctly recognized other-group faces. Whereas, higher P600 activations were recorded for participants in elderly age group while they attempted to recognize previously unseen in-group faces. This mirroring effect might reflect unique own-age face learning encoding strategies. Apart from N250 and P600, Wiese et al. did not report any P100 or N170 effects which implies that CAE requires a higher attentional load (Wiese, Komes, & Schweinberger, 2012).

1.3 Facial Aging

All faces change in a systematic way that is unique for each individual across their lifespan. During the first 20 years, the nasal and jaw regions increase dramatically in size whereas the eyes decrease relatively to rest of the face. The nose and the nasal bridge develop into a more angular shape and the forehead becomes sloppier. In adulthood, as a result of the cartilage growth the nose and the ears become larger. Wrinkles starts to appear all around the face, facial skin becomes gradually more saggy, lips become thinner and eyebrows grow larger in size (Rhodes & Anastasi, 2012).

Recently, the facial aging is discussed as a twofold process being a combined result of intrinsic (genetic e.g.) and extrinsic (environmental) factors. Both factors lead to irreversible tissue degradation, especially skin degradation. Even the bone structure remodelling occurs during individuals' lifetime. However, most of the noticeable structural changes happen in the facial soft tissue. Intrinsic skin aging is monitored by the histological level of skin layers. Sagging and wrinkling of the skin is the direct result of loss of elasticity, collagen and fatty tissue. Extrinsic factors such as ultraviolet radiation also causes wrinkling. The aging changes discussed are gender-independent however there are gender differences. The female skin is thinner, less elastic and less vascular when compared to the male skin. Moreover; in facial aging racial differences exists primarily due to the pigmentation differences (Ricanek Jr, Karl; Mahalingam & Albert, A. Midori; Bruegge, 2013).

In face images, especially for recognizing faces that have already been seen and recorded to memory, the aging factor plays a key role in facial recognition as seen from Figure 4. The changes and shifts in an individual's face might be utilized to characterize the ages of the said individual. For the purpose of facial age prediction, these facial changes might be extended to a group of individuals in order to generalize the changes for a set of individuals. As discussed before, humans possess the ability of recognizing faces from other non-face images in a master level and estimating ages from faces comes quite naturally with such recognition ability (Guo, 2013).

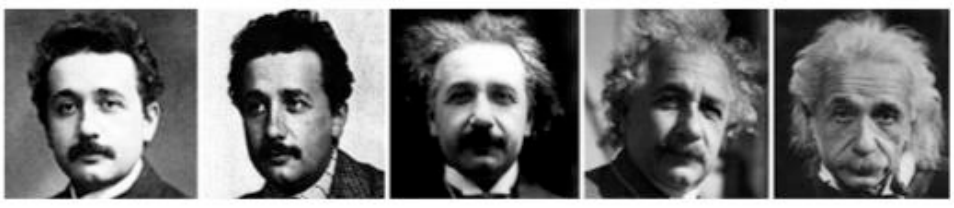

Figure 4: Facial aging of Albert Einstein through his life (Guo, 2013, pg. 232).

As the computational power and capabilities of our "computing machines" have reached a peak in our current time bin; facial aging has also been a curiosity to utilize machines to seek answers or rather seek for specialized "age rating" machines that would assist humans. Hence, the facial aging can be classified as a hot topic in recent computational vision studies. In addition to scientific curiosity; facial age estimation has many potential applications in practice as well, e.g. security, business intelligence, forensics, behaviour and emotion analysis (Guo, 2013).

Facial aging research has intrigued scientists who work in cognitive science, especially in vision. In the following table one might observe the timeline of key events occurred during facial aging research.

Table 1: Timeline of key stages in research in facial aging (Panis, Lanitis, Tsapatsoulis, & Cootes, 2015).

	Face recognition	Age estimation	Age-invariant face recognition	Age progression
Initial experimentation	1980s	1990s	2000s	1990s
Publicly available datasets	mid-1990s	2004	2004	2004
Standard performance evaluation metrics	1990s	2000s	1990s	not established
Comparative evaluations	2000	2009	N/A	N/A
Commercial systems	yes	yes	no	no
Next steps	dealing with totally unconstrained images dealing with large number of classes	dealing with totally unconstrained images/temporal information multimodal age estimation	dealing with totally unconstrained images dealing with large number of classes	systematic performance evaluation 3D age progression

1.4 Artificial Neural Networks

1.4.1 Machine Learning

In classical engineering concepts, acquisition of knowledge from the domain that is focused on is essential for each design. The problem that is focused on must be researched in detail, a mathematical model must be devised that capture the physics behind the focused subject. With the model devised, an algorithm that should solve the problem in an optimal way should be produced for high performance and accurate representation of the real-world problem. For example, designing a model that should detect the facial age of the images that is shown to the model requires a highly sophisticated knowledge on facial aging and facial aging features. And such knowledge should be should be implemented in a way that the model works as coherent as possible to the real-world age detection (Simeone, 2018).

On the other hand, as the simplest explanation, the machine learning methods replaces acquiring knowledge from the domain with acquiring a sufficiently large number of examples from the domain that is tested. These examples are traditionally named as *training set*, and such training set is fed to an algorithm which aims to produce a *machine* that is trained for to carry out the desired task. Learning happens by the choice of a group of *machines* that is selected by the learning algorithm for each training run. Such learning algorithms are typically based on the optimization of performance criteria which measures the total success rate of selected machines that are learned on the *training set* (Simeone, 2018).

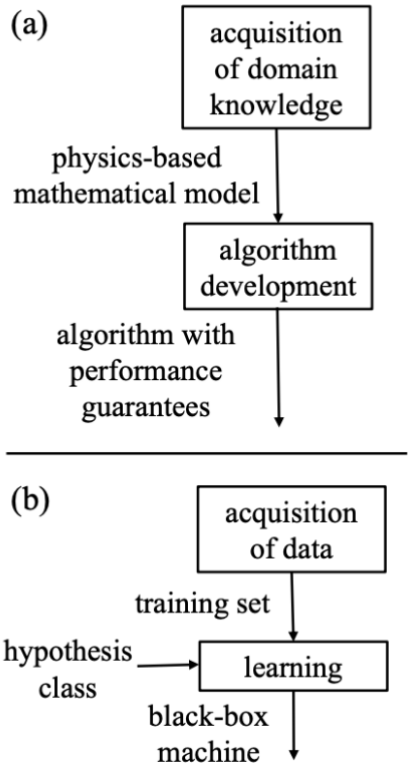

Figure 5: (a) Classical engineering problem solver design flow, (b) baseline machine learning based problem solver design flow (Simeone, 2018, p.2)

Three main classes of machine learning techniques are discussed often in the literature, *supervised, unsupervised* and *reinforced learning*. The basic details of such techniques are listed below:

- *Supervised learning:* In this method, training set are collected from pairs of input and desired output items. The goal of the *machine* is to learn a network map of connecting input items to the desired output items (Arpit et al., 2017).

- *Unsupervised learning:* In contrast to *supervised learning* this method consists of unpaired inputs, i.e. input items that are not linked to specific output items. The method typically aims to discovering the inner mechanism of the data that is generated (Hastie, Tibshirani, & Friedman, 2009).

- *Reinforced learning:* This method can be considered as a hybrid of both methods, where a supervision exists unlike *unsupervised learning,* but the supervision is not a clear linking between input-output pairs as in *supervised learning*. Rather, in *reinforced learning* the algorithm seeks for feedback from the environment after linking the pairs. With the reinforcement arriving from environment, the algorithm fulfils the learning goal (Sutton & Barto, 2018).

1.4.2 ANNs

An artificial neural network is basically a system that consists of a number of simple interconnected processors called neurons, each producing a sequence of real-valued activations. Such artificial neurons were designed to duplicate the actual biological neurons' mode of operation, such that input neurons get activated via sensors perceiving the environment. For an artificial neuron, such an environment could be a n-by-m image matrix or a matrix consisting of compressed sound. Other intermediate neurons get activated via weighted connections from previously connected neurons through input neurons to a designated output neurons (Schmidhuber, 2015).

Such systems are designed in a way that all neurons and the weighted connections in between would adjust their values in order the whole network would express a desired behaviour, such as identifying a specific feature of an image among vast number of images. Hence, ANNs mimicked object perception by establishing connections between layers of artificial neurons that have the capability of extracting the features from images provided (Cao et al., 2018).

One of the earliest mathematical single neuron models was suggested by McCulloch and Pitts, which was essentially a binary threshold unit that computed the weighted sum of input signals and with the restriction of a binary threshold, a linear discriminant or a linear classifier is finally achieved. The model responds to continuous input signals with a single binary output, with the influence of a threshold unit (McCulloch & Pitts, 1943). Those discriminating categories which are not linearly separable in the input layer required an intervention layer between input and output units. An influential solution to this problem was the backpropagation algorithm, which was made famous by Rumelhart et al. The algorithm was simply a gradient-descent solver method that iteratively adjusts the weights in order to reduce the error on output units (Kriegeskorte, 2015; Rumelhart, Hinton, & Williams, 1986).

Rumelhart et al. not only made backpropagation prominent again, they have also led to a rise of interest in neural networks in AI and cognitive science. Moreover, they have utilized backpropagation in neural network models in order to boost another solver method called parallel distributed processing (PDP) (Rumelhart & McClelland, 1986). But PDP was only manageable in basic toy-problems and when real-world problems such as vision was introduced it did not have rather influential results (Kriegeskorte, 2015).

After it was realized that PDP was not enough to solve real-world problems, neural networks fall out of prominence in 1990s. As the issue was not related to the fundamentals of the approach; but rather the computational limitations were the main

problem, thanks to the researchers such as Yann LeCun, neural networks made a comeback in 2000s.

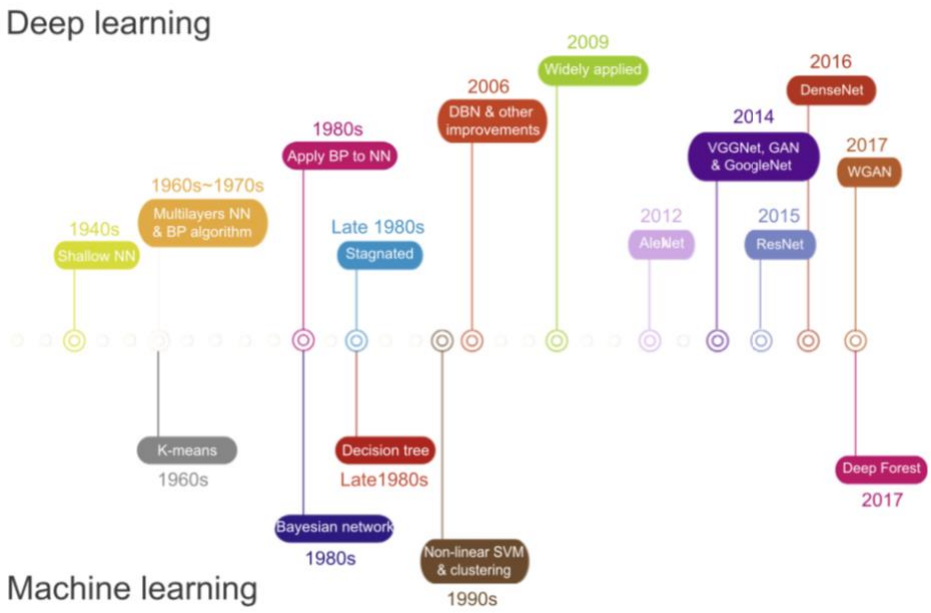

Figure 6: Timeline of the neural network and deep learning approaches. The development of deep learning and neural networks is shown in the top panel, and several commonly-used machine learning algorithms are shown in the bottom panel. NN, neural network; BP, backpropagation; DBN, deep belief network; SVM, support vector machine; AE: auto-encoder; VAE: variational AE; GAN: generative adversarial network; WGAN: Wasserstein GAN (Cao et al., 2018, p.19).

1.4.3 Deep Learning

After the comeback of neural networks and machine learning, conventional techniques were critically discussed about their limitations. Especially conventional techniques lacked processing raw natural data in a meaningful way. Constructing pattern recognizer or machine learning systems required careful engineering and tidying the natural data into a meaningful state. Moreover, considerable domain expertise was required to extracting features from raw data (for example, pixels values of an image) into a meaningful representation such as matrices etc. From that representation, the learner system or the classifier would achieve the ability to classify similar or distant patterns from the input (LeCun, Bengio, & Hinton, 2015).

Deep learning methods are representational learning methods, but in deep learning there are multiple layers of representational learning structures. Simply, the representational learning is a set of methods where a machine can be fed with raw input data and the machine would discover the representations and relations required for detection and/or classification. Such methods with multiple representational levels are obtained by forming basic yet non-linear modules that transforms those basic inputs into a higher and more abstract outputs. And as such modules are cascaded into a larger module, i.e. a deep learning network, very complex functions can be learned by the network. In classification problems, a network with higher number of representational layers increase the amount of significant input features that can be

learned, where the features generate the classifications in the first place. For example, an image data could be presented to a deep-net in the form of a two-dimensional pixel matrix and the features that are learned in the shallow layers are typically due to the presence or absence of significant edges that has particular orientations. In following layers typically analogue motifs are to be detected by spotting particular edge arrangements, even slight variations of edge positions exist. In even deeper layers, such motifs may be assembled into larger combinatory motifs which correspond to parts of features that are familiarized by the network. On subsequent deeper layers, such familiarized features' combinations may be detected as well. The most important aspect of deep learning methods it such feature layers are not designed by human hand; in contrary the feature layers are learned from the input raw data by using all-purpose learning procedures (LeCun et al., 2015).

1.4.4 Support Vector Machines

Support Vector Machines (SVMs) are first introduced by Boser, Guyon and Vapnik in their COLT-92 preceding (Boser, Guyon, & Vapnik, 1992). SVMs are classification and regression prediction tools that utilize ML to maximize prediction accuracy while overfitting to the data. Hence, they can be regarded as a subset of supervised learning machines that are used especially for classification and regression.

SVMs have become well-known when using pixel matrices as inputs; they yield higher accuracy when compared to artificial neural networks with elaborated features in a handwriting recognition task. They are also used for many applications, i.e. face recognition tasks, especially if the application includes a pattern classification task in it. Vapnik have both developed the foundations of SVMs and produced some further thought on them, such as Structural Risk Minimization (SRM) (Vapnik, 1995). In principle SRM minimizes an upper bound to the expected error and this is the key ability of SRM to have a greater notion in generalizing and categorizing (Jakkula, 2006).

1.5 Previous Results and Aim of the Study

Previous results on CAE research had their roots in CRE research since that was the first consistent effect reported in forensics and age recognition memory. Henceforth, further effects to this issue was considered and CAE investigation have born. Traditionally the mechanisms under CAE was linked with Hugenberg's Categorization-Individuation Model in which the in-group and out-group items played a key role in facial recognition tasks.

Rhodes and Anastasi's meta-analytic and theoretical review on CAE is one of the significant articles that have elucidated the phenomenon. In their analysis on previous results, they have stated that facial recognition memory is far superior for raters own age groups when compared to other age groups. Henceforth they have concluded that a significant interaction exists between the raters own age and the age of the facial image that is rated (Rhodes & Anastasi, 2012).

In our research we have several research goals. First of all, observing the CAE phenomenon on humans in order to validate it in our circumstances and to indicate that

the phenomenon is something innate and shared by most of the humankind. We have hypothesised that raters own age influences their recognition process by pulling their age guess towards their own age. More on that, our research has expanded CAE via adding an ANN based naïve rating system. In order to check ANN is over-kill for this classification problem or not, we have devised an SVM based classifier as well. Additionally, with ANN we have included a machine-based guesser whose performance can be clearly compared to the human performance and we have asked if the machine can be improved in order to overtake CAE with using any strategy that humans develop during their ordinary daily lives. Yet, our fundamental research aim on ML based experiments, especially on ANNs, was to observe and record what actions does a naïve classifier take in order to classify facial ages and moreover what does the deep-layers specialize on during training phase. At the end, the parallels between how humans conceive faces, the strategies humans have used during guessing facial aging that are reported in literature and how a classifier tries to adapt its deep-layers to categorize faces into age bins is a quest that should be achieved in Cognitive Science world.

CHAPTER 2

MATERIALS

2.1 Behavioural Experiment Equipment

Table 2: List of software used during behavioural experiment

Equipment	Company or Developer
PHP 5.6	The PHP Development Team, Zend Technologies
HTML5	Web Hypertext Application Technology Working Group (WHATWG)
MySQL	MySQL AB Oracle Corporation

Table 3: List of facial aging datasets that was considered for behavioural experiment (modified from (Minear & Park, 2004; Panis et al., 2015))

Name	Number of faces	Number of subjects	Age range	Publicly available
FG-NET-AD	1,002	82	0-67	Yes
PAML-AD	1,142	580	18-93	Yes
MORPH album 1	1,690	515	15-68	No (no longer public)
MORPH album 2	55,134	13,000	16-99	No (no longer public)
Gallagher and Chen's web collected database	28,231	28,231	0-66+	Yes
FRGC	44,278	568	teenagers and adults	Yes
Yamaha gender and age	8,000	1,600	0-93	No
Waseda database	26,222	5,320	3-85	No
Asian face database	34	17	22-61	No

Table 3 continued

Lotus Hill Research Institute	50,000	50,000	9-89	No
HOIP face database	306,600	300	15-64	No
Iranian face database	3,600	616	2-85	No
Internet aging image database	219,892	219,892	1-80	No

Table 4: Total number of participants in PAML-AD broken down by age group, race and gender (Minear & Park, 2004)

	Age							
	18-29		30-49		50-69		70-93	
Race	Male	Female	Male	Female	Male	Female	Male	Female
African-American	14	29	7	9	3	12	2	13
Caucasian	62	65	22	38	23	82	46	97
Other	38	11	0	0	2	1	0	0
Totals	114	105	29	47	28	95	48	110

Table 5: Breakdown of face images in PAML-AD by facial expression, gender and race (Minear & Park, 2004)

			RACE		
AGE GROUP	Expression	Gender	African-American	Caucasian	Other
18-29	Happy	Male	14	30	40
		Female	20	30	9
	Neutral	Male	14	62	38
		Female	29	65	11
	Profile	Male	13	43	41
		Female	16	34	9
30-49	Happy	Male	2	5	0
		Female	3	9	0
	Neutral	Male	7	22	0
		Female	9	5	0
	Profile	Male	2	9	0
		Female	3	22	0
50-69	Happy	Male	2	8	1
		Female	5	32	0
	Neutral	Male	3	23	2
		Female	12	82	1
	Profile	Male	2	10	1
		Female	7	46	1

Table 5 continued

70-93	**HAPPY**	**MALE**	**1**	**15**	**0**
		Female	9	23	0
	Neutral	Male	2	46	0
		Female	13	97	0
	Profile	Male	2	21	0
		Female	10	30	0

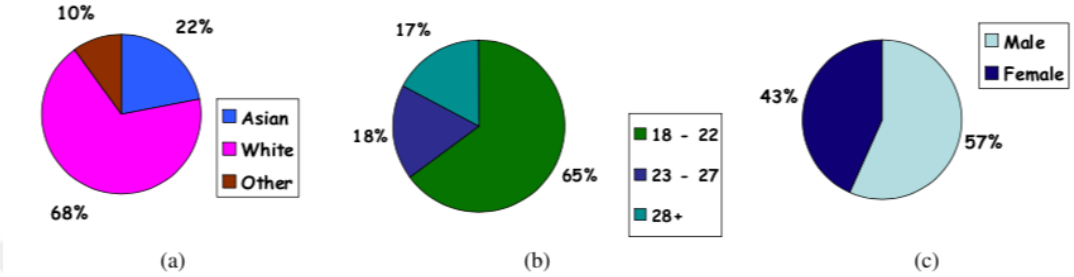

Figure 7: Demographics of FRGC ver2.0 validation partition by (a) race, (b) age, (c) sex. (Phillips et al., 2005 p.4)

2.2 ANN Based Experiment Equipment

Table 6: List of equipment utilized in ANN based experiment

Equipment	Version	Company or Developer or Manufacturer	Details
MacBook Pro	Retina 13-inch, Early 2015	Apple Inc.	Used for ANN training and testing purposes.
Personal Computer	Various	Various (Operating System: Microsoft Windows 10)	Used for ANN training and testing purposes.
ASUS RX570 GPU	-	AsusTek Computer Inc.	Used as primary GPU unit for ANN training and testing purposes.
Python	3.7	Python Software Foundation	ANN model has utilized on Python.
Keras	2.2.4	Keras Team (and various)	ANN model has constructed with using Keras.
TensorFlow	1.13	Google Brain Team	ANN model has constructed with using TensorFlow.
PlaidML	1.0	Vertex.AI (later acquired by Intel in August 2018)	ASUS GPU was utilized with the aid of PlaidML.

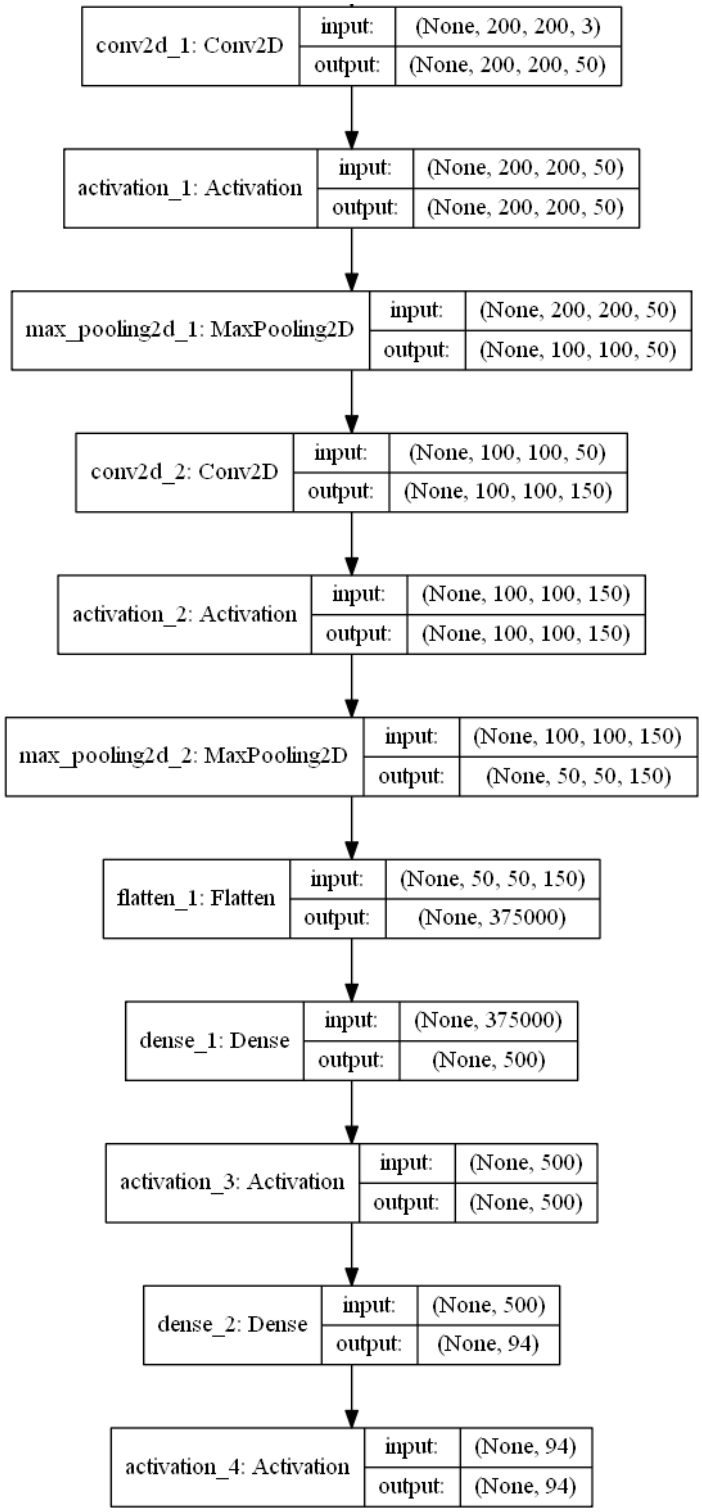

Figure 8: Detailed Artificial Neural Network design of the model that is constructed for ANN experiments. The construct was influenced from LeNet architecture (LeCun, Bottou, Bengio, & Haffner, 1998) and PyImageSearch's tutorial on image classification by NNs (Rosebrock, 2016, 2017).

CHAPTER 3

METHODS

3.1 Behavioural Experiment

3.1.1 Experimental Method and Design

One of the most crucial parts of the behavioural experiment was to select the facial aging dataset that should be the best fit to serve the testing of the initial hypothesis that was the reason of this thesis. After considering multiple facial aging datasets, for a detailed list of those datasets please refer to Table 3, it is decided that PAML-AD would be the one that suits best for the planned experiment.

The requirements that were sought from the datasets are as follows; having a diverse range of ages, gender and if possible, it should have face images of multiple races. The datasets that was not publicly available was automatically eliminated as reaching out and using those datasets was not really trivial. At the end, there were three datasets that would be beneficial for the continuum of the research; FG-NET-AD, FRGC and PAML-AD.

FRGC was the first dataset that was considered, especially due to the database having a large number of images which would be beneficial for the Deep-Net part of the experiment. But the main problem was the database was not indexed properly for facial aging, rather their intention was to design a dataset to test the capabilities of the facial recognition at the time of the design (Phillips et al., 2005). Moreover, as seen from Figure 7, FRGC's race and sex partition was more than satisfactory but the age range is rather restricted; 18-28+. Therefore, due to the lack of proper age range, FRGC was eliminated.

FGNET-AD was the second dataset that was considered for behavioural experiment, and a demo experiment was designed on top of it. The age range and number of images that FGNET-AD had was satisfactory especially when compared to FRGC. The database consists of a total of 1002 portrait photographs of 82 people which are taken in different ages during their lifetime. The maximum age of faces was 67, minimum age was 0 (infant baby), mean age was 25,43, median age was 22 and the standard deviation was 17,05.

Even though the age range and number of facial images looked satisfactory enough, main problem of FGNET-AD was its lack of uniqueness of the facial images. The dataset had 1002 images, but all of these images belonged to only 82 people. Hence; to eliminate the familiarity effect that can arise during behavioural and computational experiments, FGNET-AD had eliminated as well.

PAML at University of Texas Dallas was constructed the PAML-AD in order to answer the rising demand on facial imaging and facial aging research. (Minear & Park,

2004). PAML did not hesitate to share their dataset and the dataset was satisfactory enough for both experiments' purposes. PAML's database includes various sub-databases such as "happy faces", "neutral faces", "profile faces". The main sub-database that is used for our scope is the one with the "neutral faces" as it would perfectly suit both of the experiments hence facial emotion was not considered and even omitted during research.

The experimental design was aimed to stimulate the participants' facial age recognition mechanism, but in an isolated setting. The setting was decided to be an isolated one, as even the facial age, gender and race themselves interact each other and result in interference (Li & Tse, 2016). Since this phenomenon is well-reported, any variables other than facial age, gender and race was intended to be eliminated.

During the demo experiment conducted on top of FG-NET-AD it was realized that inviting participants one by one to an isolated setting and saving their individual answers was not really optimal for a behavioural experiment that would require a high number of participants to validate the initial hypothesis. To compare with the final design, the demo was held with using a terminal console running on the same Apple MacBook Pro 13 Inch with the oral instructions of the conductor. All participants entered their age guesses on same computer one by one and the results were saved locally. With this experimental setup, only 15 participants agreed to join, and 23 runs were saved from 15 participants.

Moreover, FG-NET-AD was a face image dataset which was purely constructed to investigate the facial aging. Since our hypothesis is linked more with facial age, rather than facial aging; PAML-AD was decided to be used as the main dataset for both behavioural and computational experiment.

In order to increase the participation and make the experiment easy-to-access for all participants, the experiment had moved to a website[1]. A web-based application was decided since there are not many significant extrinsic factors that would drastically change the ratings of participants. Moreover, inviting a high number of participants to a laboratory for such a trivial task would be a bottleneck for the experiment and serious schedule conflicts would have arisen for participants.

3.1.2 Sample

A total of 367 participants were accepted to the experiment. From 367 participants, 250 of them had signed the consent form that was required to process their experimental data. From 250 participants that have signed the consent form, 54 of them had made significant errors during their run and the errors made cannot be fixed by hand. Hence in order to prevent any further confusion, the faulty data was removed from the database. Finally, 206 unique participants' experimental data was acceptable for the behavioural experiment and they were saved to the final database.

[1] Experimental setup can be observed from: http://cagataygursoy.xyz/test_index.php

In order to match the actual age range, obtaining a wide range of participants were key for this experiment. Therefore, having a relatively high number of participants from different age groups matched the initial aim for the participant database.

The age distribution of the participants can be observed from Figure 9. As mentioned before; number of valid participants was 20, the mean age was 38,89 and the standard deviation was 13,60. Even though there were two major clusters around 20-25 and 50-55 bins the dispersion of the ages seems to be satisfactory in order to have an accurate attempt to validate the hypothesis.

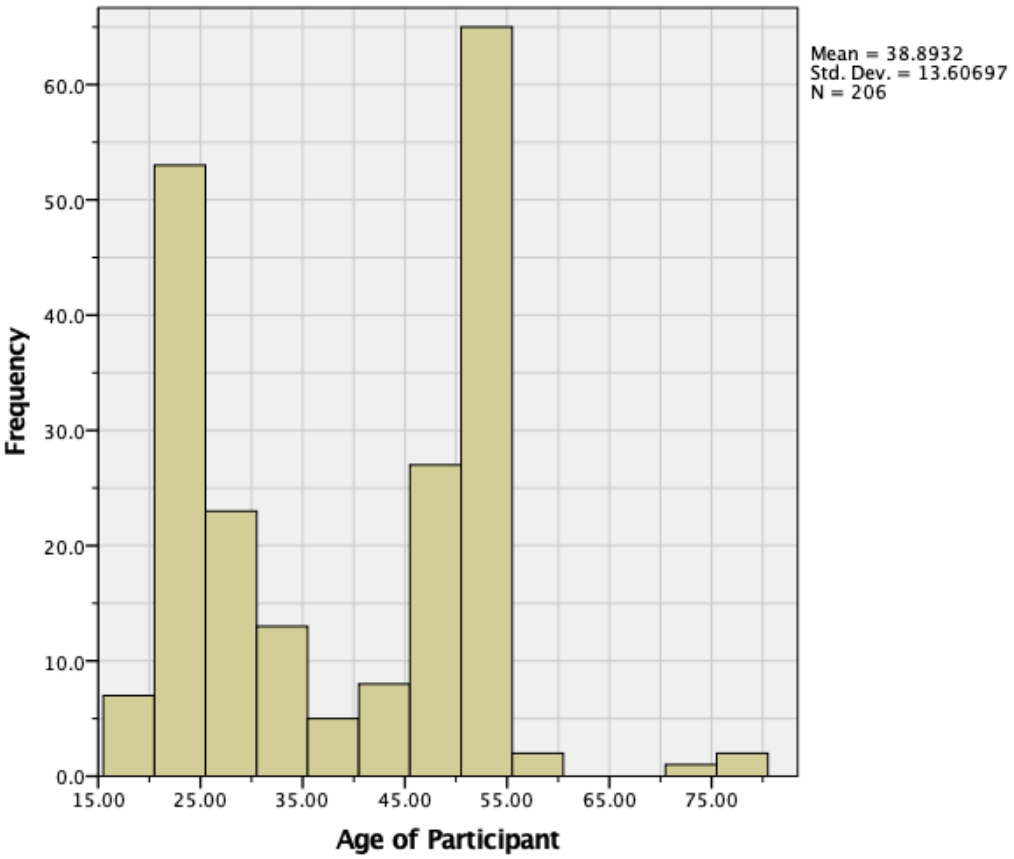

Figure 9: Histogram of participants' age.

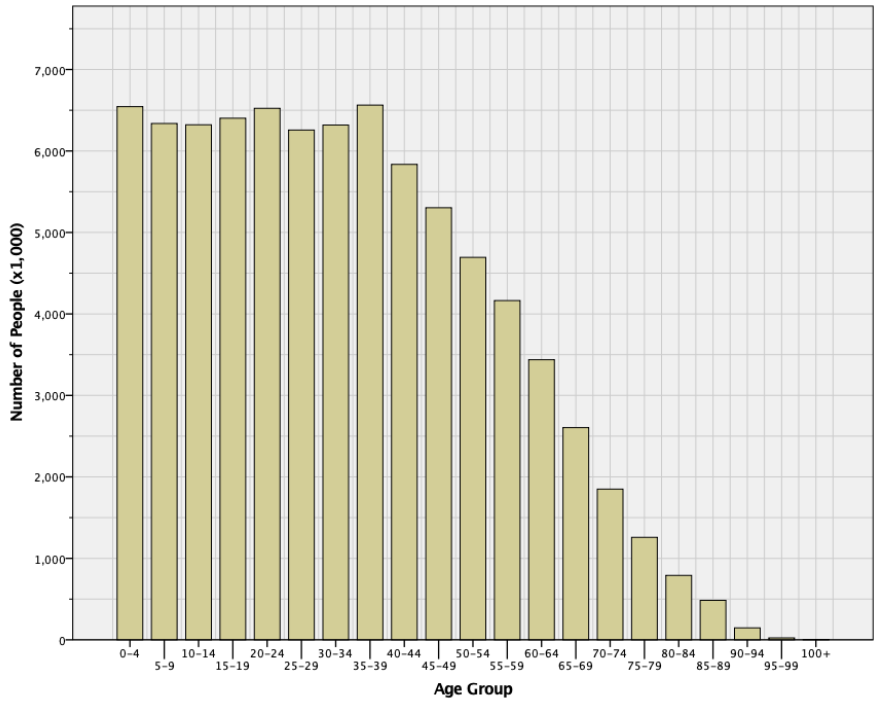

Figure 10: Turkish age distribution segmented by age groups. (Turkish Statistical Institute, 2018a)

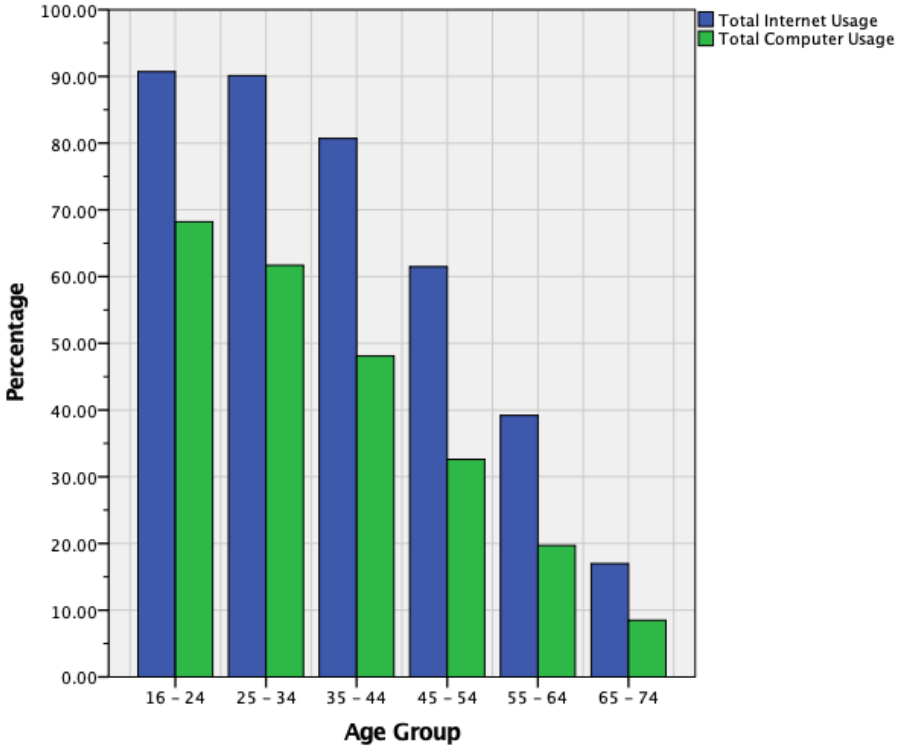

Figure 11: Internet usage distribution among age groups in Turkey. (Turkish Statistical Institute, 2018b)

If Figure 9 and Figure 10 compared regarding accuracy, it seems like our participants were not an exact match with the actual Turkish age distribution. To understand this

situation, internet usage data for Turkey is referred as the experiment was primarily advertised by means of internet and social media. So, when Figure 10 and Figure 11 were to be interpolated, the peak at around 20-25 age bin is clearly acknowledged as internet usage at 20-25 age bin is clearly high. Yet, the peak at 50-55 age bin cannot be acknowledged by Figure 10 and Figure 11. Thus, it can be concluded that the experiment had reached people at 50-55 age bin through internet and social media more than any other age bin.

In order to investigate a presumptive relation between gender and facial age prediction, gender information was collected from participants as well. In Figure 12 one can observe the distribution of genders among participants. Moreover, as seen from Table 7, from 206 participants 90 of them were female and 116 of them were male.

When compared to Turkish gender distribution, the experiment's gender distribution is acceptable. From Table 7 and Table 8 it can be observed that the experiment's gender percentage is 43,7% versus 56,3% (Female versus Male) whereas the actual gender distribution from TurkStat data is 49,8% versus 50,2% (Female versus Male). Even though the actual distribution percentages are significantly closer to each other, our distribution looks somewhat acceptable, if not completely realistic.

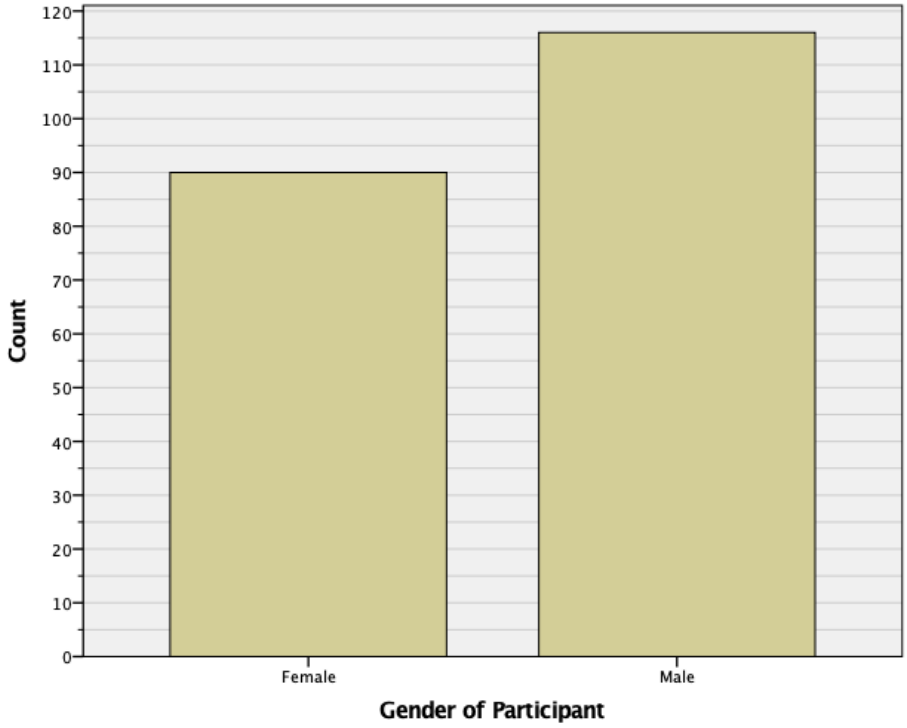

Figure 12: Histogram of participants' gender.

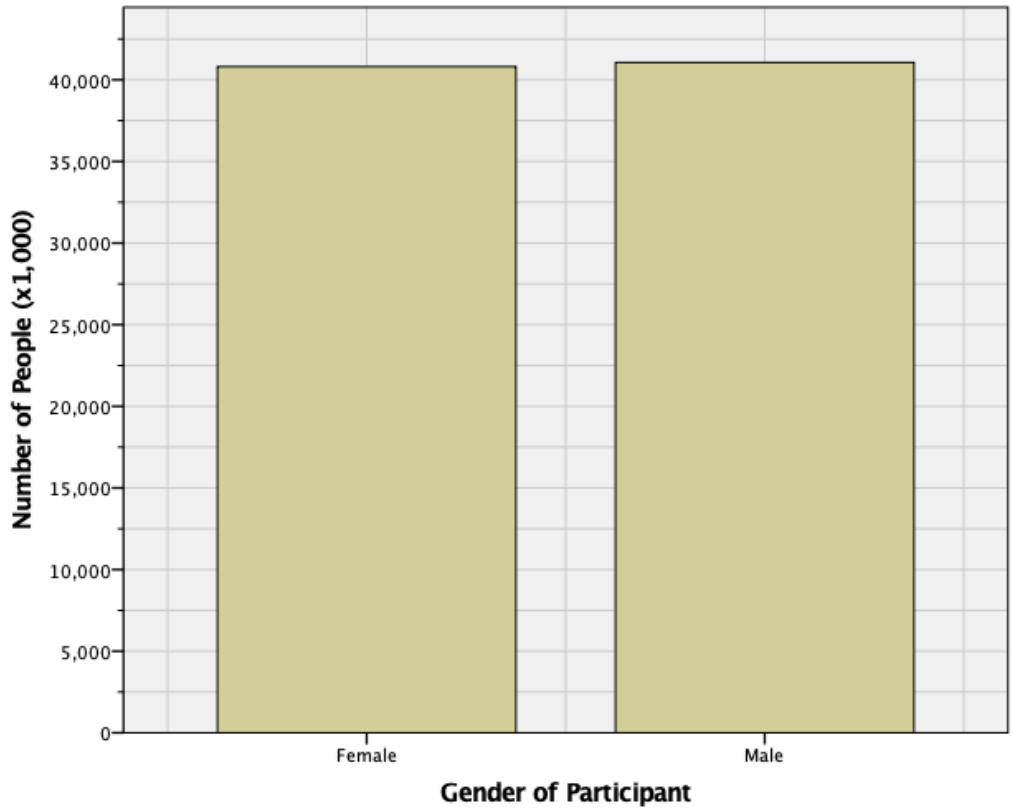

Figure 13: Turkish gender distribution. (Turkish Statistical Institute, 2018a)

Table 7: Frequencies of participants' gender.

Frequencies of Participant Genders

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	90	43.7	43.7	43.7
	Male	116	56.3	56.3	100.0
	Total	206	100.0	100.0	

Table 8: Turkish gender distribution frequencies. (Turkish Statistical Institute, 2018a)

Frequencies of Turkish Gender Distribution

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	40.808.147	49.8	49.8	49.8
	Male	41.059.075	50.2	50.2	100.0
	Total	81.867.222	100.0	100.0	

3.1.3 Stimuli

Couple of selected face images from PAML-AD's original image set can be observed in Figure 14. The faces on the images were all neutral faces, and the background of

the image is the same grey background for each image. The images in PAML-AD were collected in site visits which were arranged with the aid of two college student unions, a shopping mall and two senior citizen festivals in northern Ohio and southern Michigan. The experimenters from PAML who collected facial images had explained their purpose of visit for those all approached their table and agreed to participate in their research. Moreover, the participants have signed a consent form in which PAML asked for their permission to use their facial image data for research purposes. For each participant sex, age and race/ethnic background were recorded but participants' names were excluded from any recorded data. The participants who have agreed upon the consent form have asked to stand in front of a neutral grey background where one to three photographs were taken. For each participant, a neutral posed picture is taken; but for some of the participants who have agreed to continue, a happily posed and a right-faced profile picture were taken as well. The pictures were saved in a 640x480 pixel resolution and in bitmap format. (Minear & Park, 2004)

Figure 14: Randomly selected faces from original PAML-AD image set.

As the original images have larger grey backgrounds, the risk of losing attention for participants were considered. Hence the images were cropped in order to make faces in the images stand out, even though the background was always a plain grey one. In Figure 14 one can observe four randomly selected face images from the dataset which are uncropped (640x480px). In contrast, Figure 15 depicts the same four images but this time the images were cropped (351x480px) to emphasize more on faces.

Figure 15: Cropped images from PAML-AD image set.

3.1.4 Procedure

The participants have asked to visit the experiment's website and before starting the experiment they were requested to sign the consent form. An example blank consent form can be observed in appendix. Every participant had to read and agree upon the consent form in order us the process their data.

After a participant has signed the consent form, they were redirected to the initial page to enter their name, age and gender information. Even though we did not process their name data, their names were requested in order to differentiate the runs since it was not prohibited for them to participate to the experiment multiple times.

After the participants have entered their personal information; they have seen 20 random face images, one by one. As they have seen the faces, they were requested to input their age guesses to a textbox labelled as "Your Age Guess". As there were no time restriction, no new face would appear on the screen until the participants have entered their guess. After their submission, next face image has presented to them and until the end of their run their guesses will be collected one by one.

At any time during the experiment's run, every participant is allowed to terminate their run by just closing their browser's tab where our experiment was active. In such cases, their data will be deduced from our final data as the experiment will not be considered as complete.

Finally; when a participant has seen and rated all 20 random faces, they have redirected to the final page of the experiment. From this page, they can gather further information about our research and CAE. At this point, a participant was completed their participation and they are free to leave the webpage or start a new run of the experiment.

3.2 Artificial Neural Network Based Experiment

3.2.1 Experimental Method and Design

As the selection process of face image dataset was already lengthily discussed in Behavioural Experiment section, it will not be further discussed in this section. Instead, the structure of the ANN will be discussed in a detailed manner.

For the architecture of our research, Yann LeCun's LeNet neural network architecture was used as the basis. In his influential published research, it has stated that with appropriate network architecture, gradient-based learning algorithms can classify high-dimensional patterns. LeCun et al.'s work involved classifying hand-written digits and the capabilities of LeNet is clearly shown in their research. Simply put, LeNet is a Convolutional Neural Network (CNN) architecture that is capable of classifying 2D images with utmost precision and it was shown that LeNet can outperform other architectures especially in accuracy and memory usage (LeCun et al., 1998).

LeNet was initially constructed to classify handwritten digits and characters, especially for Optical Character Recognition (OCR). But with slight modifications, LeNet can be converted into a powerful face recognition tool, as shown in Lin et al.'s research (Lin, Cai, Lin, & Ji, 2016). With using their research and a tutorial that was published in PyImageSearch website, our ANN architecture is constructed (Rosebrock, 2017). A basic scheme of our architecture is shown in the following figure.

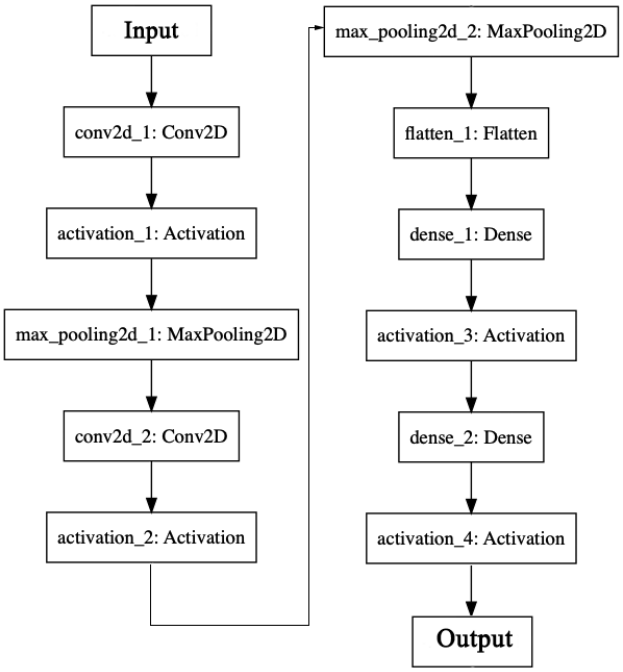

Figure 16: ANN architecture of the experiment. A detailed version is given in Figure 8.

Table 9: ANN structure and details about layers. Visualization is given in Figure 16.

Layer No	Layer name	Input and output dimensions of the layer	Working principle and details of the layer
1	Conv2d_1	Input: 200x200x3 Output: 200x200x50	First 2D convolutional layer. Basically, what convolutional layer does is to check if any patterns exist in the input
2	Activation_1	Input: 200x200x50 Output: 200x200x50	This activation layer is performed by using a Rectified Linear Unit (ReLU) activation function. ReLU
3	Max_pooling2d_1	Input: 200x200x50 Output: 100x100x50	What pooling layer does is to split data into sub-regions and return the maximum value of the sub-regions.
4	Conv2d_2	Input: 100x100x50 Output: 100x100x150	Second convolutional layer. Same working principle as the first one.
5	Activation_2	Input: 100x100x150 Output: 100x100x150	Again, this activation layer is performed by ReLU function.

Table 9 continued

6	Max_pooling2d_2	Input: 100x100x150 Output: 50x50x150	Second pooling layer.
7	Flatten_1	Input: 50x50x150 Output: 1x375000	Flattening layer. Basically, this layer converts three-dimensional image matrices into one-dimension vectors.
8	Dense_1	Input: 1x375000 Output: 1x500	Dense layer is simply a fully connected layer.
9	Activation_3	Input: 1x500 Output: 1x500	Third activation layer, again ReLU governs.
10	Dense_2	Input: 1x500 Output: 1x94	Second dense layer.
11	Activation_4	Input: 1x94 Output: 1x94	Final activation layer, this time in order to get the age guesses Softmax function is used.

Basically, our architecture does two principal operations; firstly, it works on 2D image matrices and then it "flattens" those matrices into 1D vectors to classify the image's age. 2D image that is fed to the ANN will be processed first by two Convolutional-ReLU-Pooling layer triplets. The outputs of those two triplets was planned to give images. In those images, it was expected that facial features that were helping ANN to classify the image were to be implied by higher levels of activations. Then after flattening the partly-classified 2D image matrices into 1D vectors it was planned to search for the highest level of activations which would give the age prediction of ANN.

For train and testing purposes, initially a MacBook Pro's CPU was used. As expected even for smallest images, e.g. 20x20px images, the training and testing sequences took several hours to complete. Hence, the architecture was constructed in a desktop computer which was running on Windows 10 and has an AMD Radeon GPU in order to increase timing performance. But it was realized that even though Python's Keras library can be run on GPUs, the GPU must be a Nvidia one. So, to overcome this, another Python library called PlaidML is utilized. In short, PlaidML is a library that

allows ANNs that was constructed Keras to run on GPUs that are not Nvidia. PlaidML was developed by Vertex.AI, which was acquired by Intel in summer 2018[2].

3.2.2 Sample

A total of 580 facial images which had neutral expressions were gathered from PAML-AD to construct the dataset of the ANN. Detailed breakdown of the facial images can be observed in Table 10.

Table 10: Total number of facial images in PAML-AD broken down by age group, race and gender (Minear & Park, 2004).

	Age							
	18-29		**30-49**		**50-69**		**70-93**	
Race / Gender	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
African-American	17	29	7	9	4	13	2	13
Caucasian	83	72	22	38	29	77	46	97
Other	18	4	0	0	0	0	0	0
Totals	118	105	29	47	33	90	48	110

3.2.3 Stimuli

As mentioned in behavioural experiment, we have further cropped the already cropped facial images before showing to human participants. Now at this stage, because of performance issues, another cropping is required. In order ANN to process the images, they were required to converted into to square matrices. Moreover; due to limited GPU memory capacity, 351x480px images were converted into 200x200px images before feeding them to the ANN.

Figure 17: Transformation of facial images. Leftmost image was original facial image provided by PAML. Middle one was the cropped image that was shown to participants in behavioural experiment. Rightmost image was the image that was presented to ANN for the experiment.

During training phase 60% of the images were split into "training dataset", 20% of the images were split as "validation dataset" and remaining 20% were split as "testing dataset" to achieve the highest performance.

[2] Further information about PlaidML can be found in https://github.com/plaidml/plaidml

3.2.4 Procedure

ANNs were trained with utilizing Python's Keras library. As mentioned before Keras is a highly flexible DL framework, which can be used together with other DL frameworks, Python packages and DL languages such as TensorFlow and PlaidML. An AMD Radeon GPU was utilized together with PlaidML for training and testing phases to achieve the most optimal timing.

200x200px face images were saved to 3D RGB encoded matrices at first to train a model, after training process same images were used to test the ANN model. Several methods used to further test the constructed model's classification capabilities:

3.2.4.1 All PAML-AD images trained and tested

All 580 face images were used to train and test the model. Firstly, each image was grouped with respect to their own age. As the ages of the faces were varied between 18-93, 75 age group folders were constructed, and images were copied to their respective age folders. In ANN model, 94 age groups (age range 0-93) created in order to test if the model predicts any page out of the actual age range.

When image dataset is split in 60%-20%-20% training-testing-validation ratio, we had 348 images for training and 116 images for test dataset and again 116 images for validation.

3.2.4.2 Trained with using 10-adjusted groups

For this run; in order to normalize the frequencies of facial ages, using Scott's Rule in Equation 1, an optimal bin size is calculated.

$$h = \frac{3.5\hat{\sigma}}{n^{1/3}} \tag{1}$$

With using Scott's Rule, the optimal bin size is ~7 for our case. Hence, the images were split up with bins sized 7 age difference. In Figure 18 one might observe the histogram of grouped images.

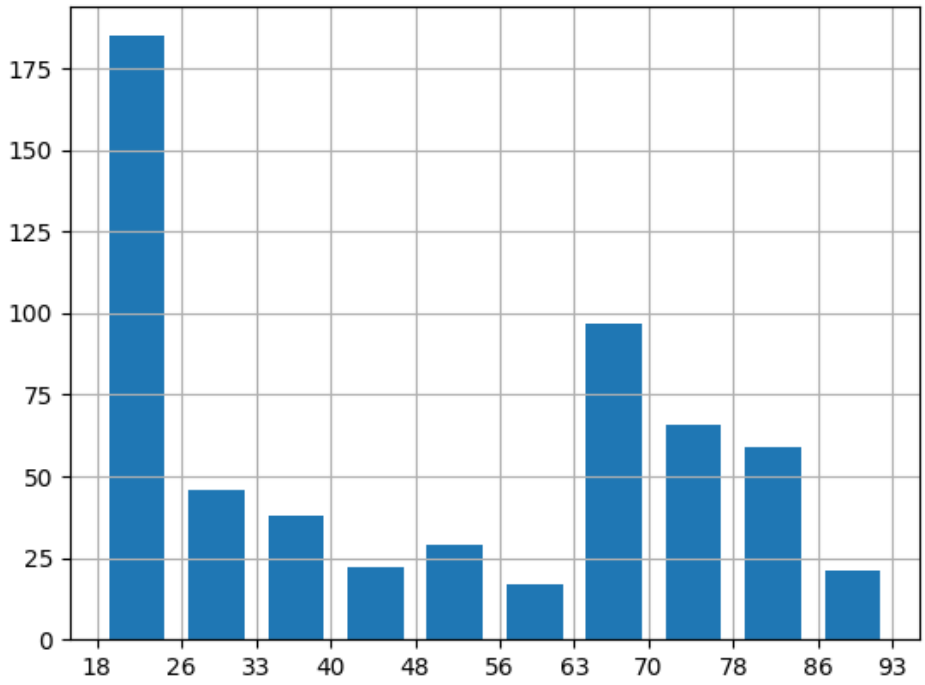

Figure 18: Histogram of adjusted age groups with respect to the bin size calculated via Scott's Rule (bin size = 7).

Lowest number of faces were in 56-63 bin, which was 17. So, images were randomly assigned to all ten groups with the restriction of exactly 17 images for each age range bin. After this adjustment, the model is trained with the images that is adjusted.

When image dataset is split in 60%-20%-20% training-testing-validation ratio, we had 136 images for training and 34 images for test dataset and again 34 images for validation.

After the training all images were tested for model's age prediction.

3.2.4.3 Trained with using 10-adjusted groups, only selected images tested

In this run, the model that is created for the Section 3.2.4.1 is used. But the difference was only the images that were randomly chosen for training was tested instead of testing all 580 face images.

3.2.4.4 Trained with using 5-adjusted groups

For this run, a similar setting to Section 3.2.4.1 was constructed. Only difference was this time the bin size was hand-picked, and it was 15. In Figure 19 one might observe the histogram of grouped images.

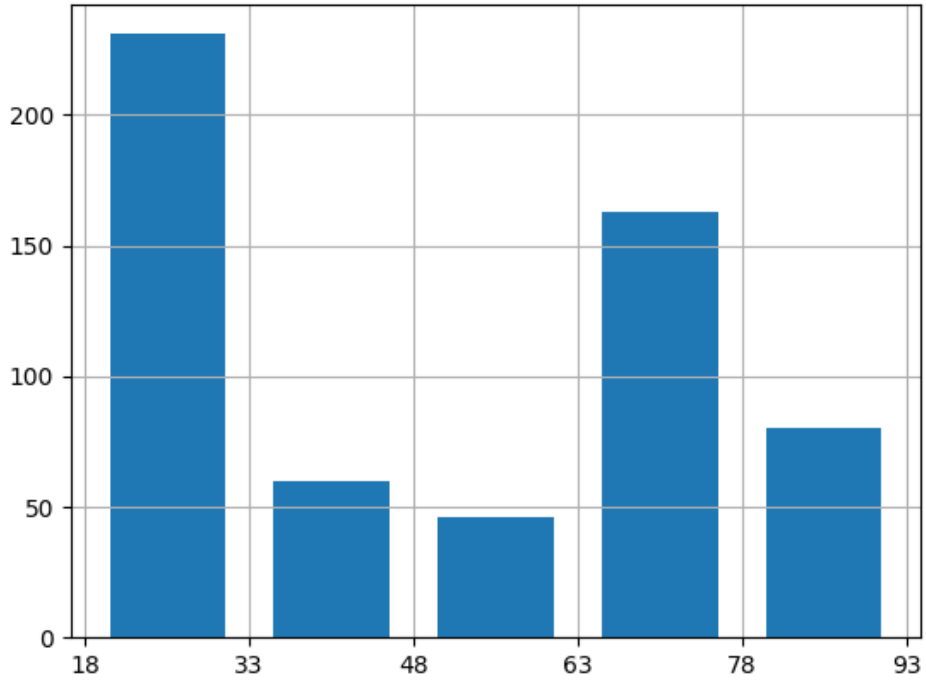

Figure 19: Histogram of adjusted age groups with respect to the bin size picked by hand (bin size = 15).

Lowest number of faces were in 48-63 bin, which was 46. So, images were randomly assigned to all ten groups with the restriction of exactly 46 images for each age range bin. After this adjustment, the model is trained with the images that is adjusted.

When image dataset is split in 60%-20%-20% training-testing-validation ratio, we had 184 images for training and 46 images for test dataset and again 46 images for validation.

3.2.4.5 Trained with using 5-adjusted groups, only selected images tested

In this run, the model that is created for the Section 3.2.4.4 is used. But the difference was only the images that were randomly chosen for training was tested instead of testing all 580 face images.

3.2.4.6 Trained with PAML-AD, tested with FGNET-AD

This run was specifically constructed in order to test the capabilities of our model when completely unknown faces were presented to it.

The model was trained with PAML-AD's all 580 neutral face images. The training method was identical to the method in Section 3.2.4.1. But instead of PAML-AD's own face images, FGNET-AD's face images were presented to the model and the

predictions were saved. In order to fit our purposes, FGNET-AD's images were resized to 200x200px images as well.

3.2.4.7 Trained with PAML-AD's neutral faces, tested with PAML-AD's happy faces

After completing a run on faces that the model has not seen, another test is applied to the model. This time, PAML-AD's own faces were used but test faces was not neutral faces, rather they were faces which had happy expression. Rest of the procedure was identical to Section 3.2.4.1's procedure and at the end of the run the age predictions of the model were saved.

3.3 SVM Experiment

3.3.1 Experimental Method and Design

In order to validate ANN's performance was not over performing for the given classification task, an SVM based model was designed as well.

To match up the features of ANN; SVM model is again constructed in Python, using Scikit-learn library. For outputs; receiver operating characteristics, confusion matrix and Scikit-learn's classification report were prepared.

3.3.2 Sample

As the same images were used in this experiment, the details are identical with the sample details given in Section 3.2.2.

3.3.3 Stimuli

Stimuli were the same, henceforth the details are given in Section 3.2.3.

3.3.4 Procedure

Just like the ANN experiment, all 580 face images were used as training and testing sets. Yet, unlike ANN experiment, this time 80% of the images were used as training set and 20% of the images were reserved as testing set.

200x200px face images were saved to 3D RGB encoded matrices at first, then those matrices were converted into 1D vectors in order to feed it to the SVM. SVM classifier's kernel was chosen as linear, in order not to turn classifier into a more-than-enough and too powerful tool; since the existence of this experiment is to solely check if ANN is too powerful.

Moreover, in order to test further, image datasets that are edited for 5-adjusted groups and 10-adjusted groups ANN experiments are used for SVM datasets as well. Since the dataset details are identical to ANN experiment, it is not discussed here. So, for details please refer to Section 3.2.4 and respective subtitles, namely Section 3.2.4.2 and Section 3.2.4.4.

CHAPTER 4

RESULTS

4.1 Behavioural Experiment Results

A total number of 4356 ratings were obtained from 206 participants who agreed to join the online experiment. During the experiment participants tried to guess the age of 582 facial images. The real ages of the people in the image database ranged from 18 to 93, which is summarized in the histogram in Figure 20. The age distribution of the people in our picture database was not uniform, where people in the age range 20-30 were the most frequent, followed by 60-80.

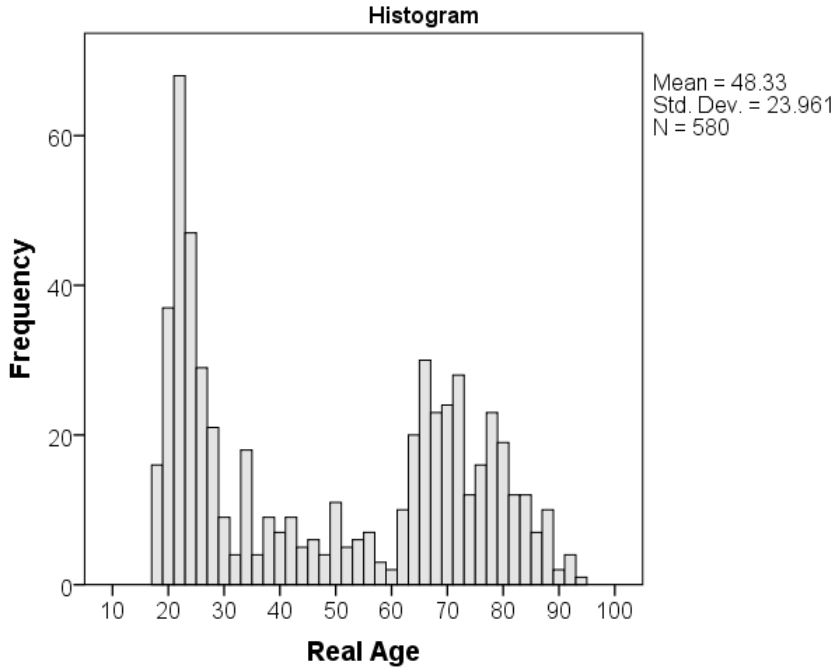

Figure 20: Age distribution histogram of images in PAML-AD.

The 206 participants rated on average 20 pictures sampled from the database. Each facial image was rated on average by 7 different participants. Figure 21 shows the histogram and Figure 22 shows the boxplot for the errors made by human raters when they guessed the age of the person in the photographs. The errors exhibited a symmetric distribution around 0 (M =0.81, SD=8.20), mostly within +/- 10 years. So, participants of the study were approximately accurate in predicting the age of the persons in our image database. Due to the non-zero kurtosis (.73), a Kolmogorov-Smirnov test showed that the error distribution deviated from a normal distribution, $D(4356)=.064$, $p<.01$.

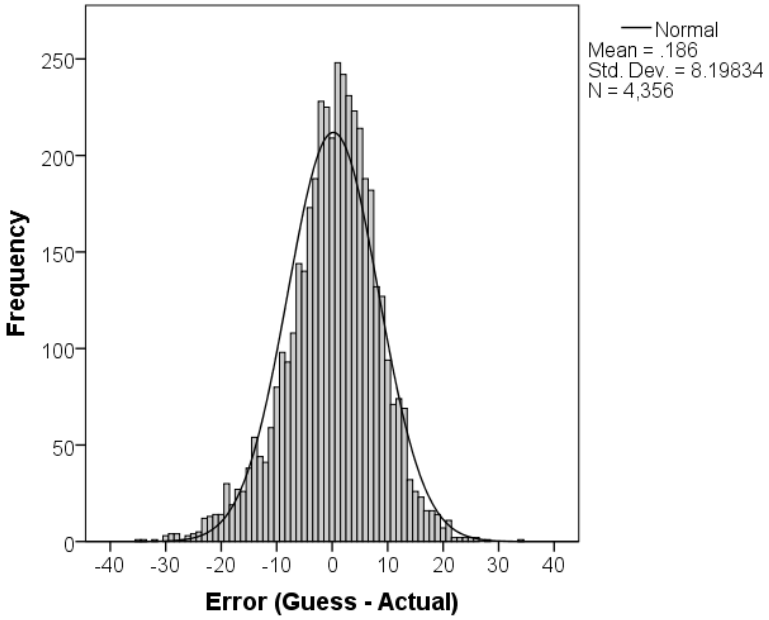

Figure 21: Histogram of the errors done by human raters while guessing the age of the image shown.

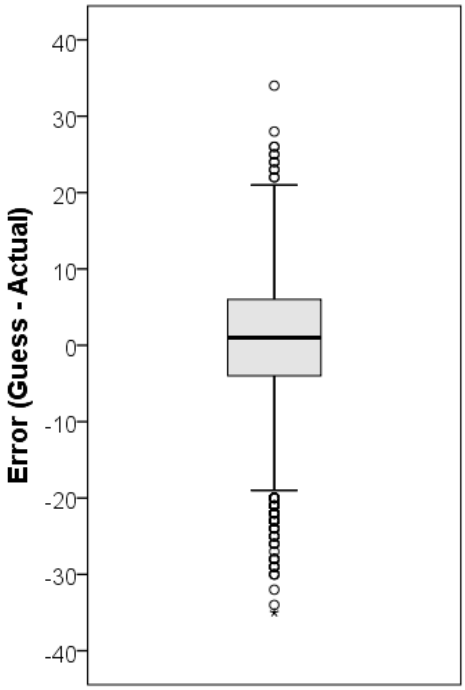

Figure 22: Boxplot of the errors done by human raters while guessing the age of the image shown.

Although the margin of error is within +/- 10 years, there is a strong correlation between the participants' guesses and the real ages of the people displayed in the images, r=.94, p<.001. A simple linear regression model where guessed age is the predictor and the real age is the outcome variable can account for %88 of the variance.

Following scatterplot below shows the relationship between the real and the guessed ages.

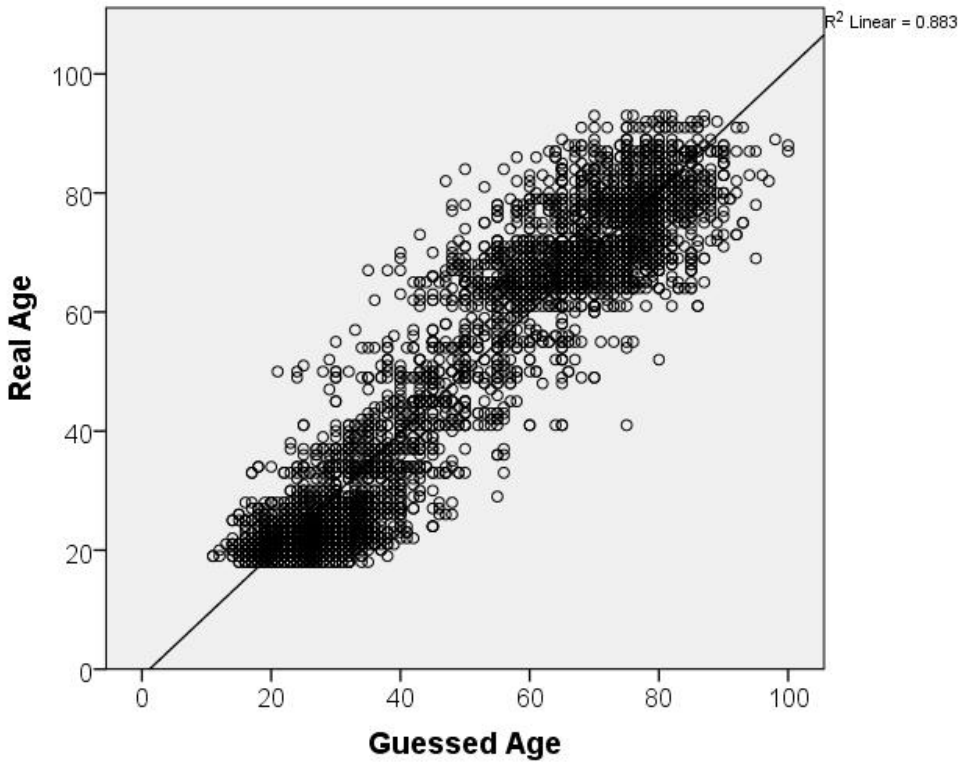

Figure 23: Scatterplot showing the relationship between real age of the face images and the age guesses of the raters.

We also investigated whether there is a pattern in the distribution of the errors as a function of the age distance between the rater and the person in the picture. In related literature it was reported that people tend to make better predictions for people who are about the same or lower age than they are (Tanaka & Pierce, 2009). One possible explanation for this tendency is that people who are at a certain age have a first-hand experience of the ages that they have lived through, so they may judge better for an age range lower than their own age.

In order to check whether our data set supports this view, we computed the difference between the age of the participant and the real age of the person in the image (i.e. the objective age distance), and the age of the participant and the guessed age of the person in the image (i.e. the perceived age distance). By this way we aimed to concentrate the data around the ratter's age, which is represented as zero. The error is defined as the difference between the guessed age and the real age of the person in the image.

Correlation analysis suggested that there is a positive relationship between error and objective distance, $r=.38$, $p<.01$, whereas there is almost no correlation between error and perceived distance, $r=.09$, $p<.01$. The scatterplots in Figure 24 and Figure 25 did not immediately present an obvious funnel like picture where the error range decreases around and above zero, which represents the target age range where we expected the participants to get more accurate in their estimations.

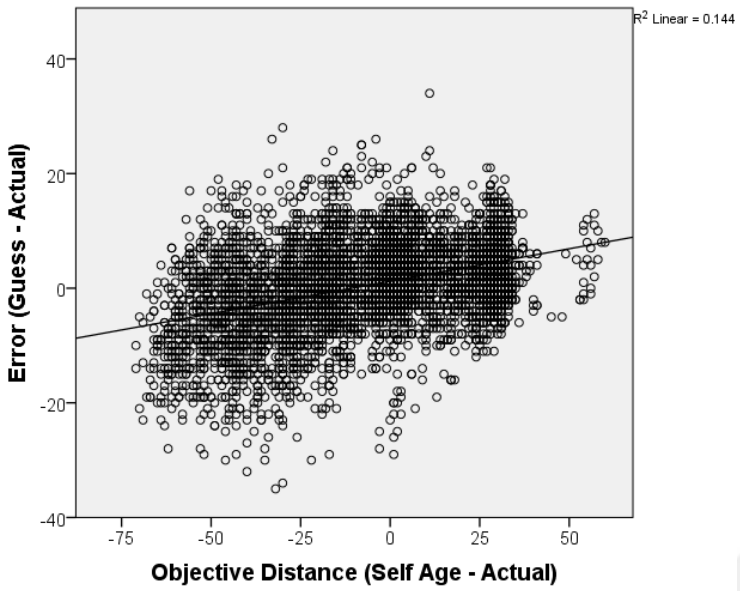

Figure 24: Scatterplot of error versus objective distance. Objective distance is calculated as difference between the age of the rater and the age of the image seen.

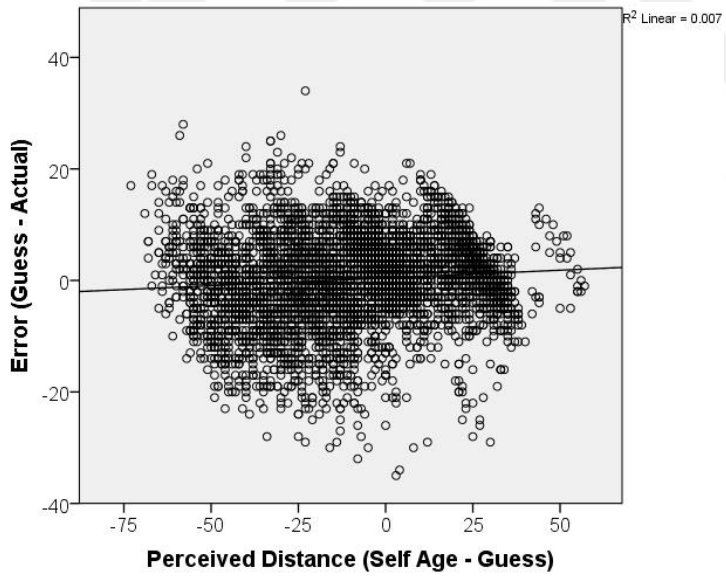

Figure 25: Scatterplot of error versus perceived distance. Perceived distance is calculated as difference between the age of the rater and the raters' age guess for the image seen.

However, a strong positive correlation was observed between objective and perceived error, $r=.96$, $p<.01$. The scatterplot in Figure 26 also indicates a slight funnel pattern where the variability is reduced for data points starting around the center 0 to positive differences (i.e. for pictures with younger people). This suggests that the participants' age judgments were consistent with respect to their perceived age difference with the person in the image.

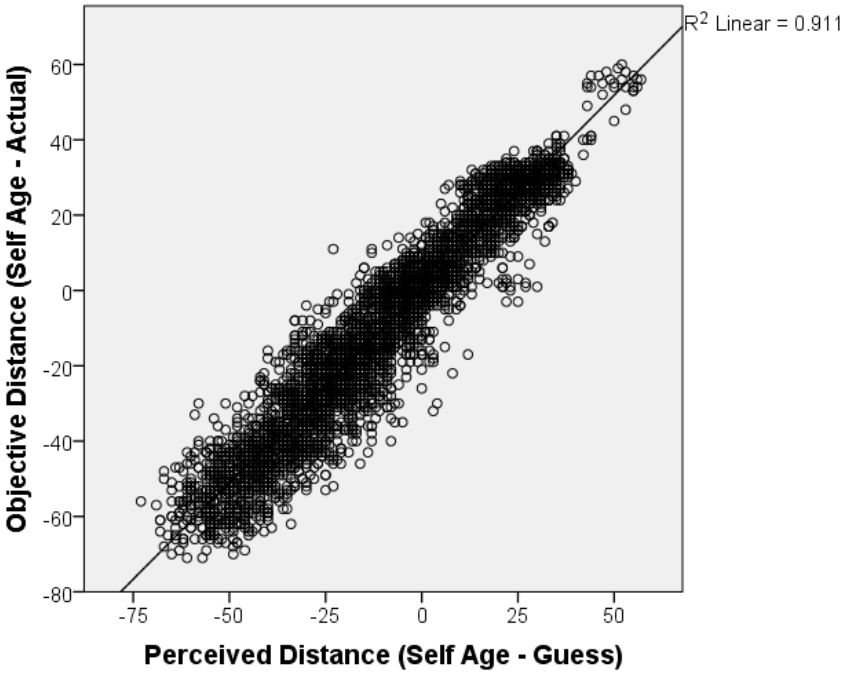

Figure 26: Scatterplot of objective age distance versus perceived age distance.

To test this effect further, we split the data into 6 groups along the objective distance dimension by considering standard deviations as thresholds. We used +/- 2 standard deviations around the mean value of objective distance (M=-9.18, SD=27.43) to bin the data along 6 intervals. The means plot in Figure 27 shows the average error observed for each interval. Since the homogeneity of variance could not be met, a Brown-Forsythe corrected one-way ANOVA was used to compare the intervals in terms of their mean difference. The test showed a significant difference among the intervals in terms of observed average error, $F(5, 608.2) = 178.34$, $p<.001$, partial $\eta 2 = .15$. Bonferroni corrected post-hoc tests suggested that the difference between all pairs of intervals were significant except the last two intervals (19,45) and >46. The error is closest to zero for the intervals where the objective age difference between the rater and the person in the image is small. The overall distribution of error in the 6 intervals show a sigmoid like pattern that gets closer to zero error in the middle intervals. This seem to suggest the estimations were better for the age range close to the raters' age.

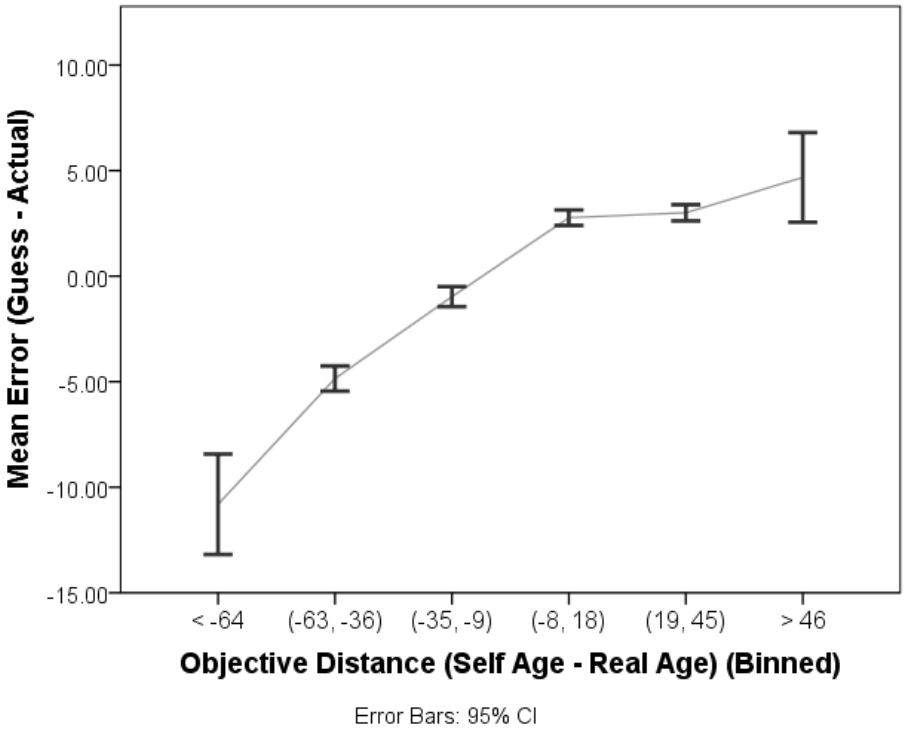

Figure 27: Average error observed in six defined groups.

4.2 Artificial Neural Network Based Experiment Results

For each experimental setting, respective results are reported in their own headings in the following sections. In the last section, all the experimental settings' results are reported in a comparative manner.

4.2.1 All PAML-AD images trained and tested

Training accuracies of the constructed ANN model is reported in Figure 29. It is obvious that with respect to training accuracy, our model meets our initial expectations. The model's training accuracy was around 90%, which is more than acceptable for a classifier with 94 groups.

Even though the training accuracy was around 90%, the validation accuracy has never exceeded 20%. There are various reasons for this, and most of them were due to the size constraint of our database. As PAML-AD has a limited number of faces, distribution of the face ages can be observed in Figure 28, validation accuracy was significantly low. This is mainly due to having a dataset with not enough data points. This causes model to have an incline to predict face ages into classes that have large number of face images more frequently than classes that have lower number of face images.

To overcome this problem the validation dataset size is increased, which was initially 20%. This increment in validation dataset size did not yield a significant result even though this boosted validation accuracy from 15-20% to almost 40%, training accuracy fell significantly from 90-95% to 75-80% and it caused instability across both

validation and training accuracies. Regarding result graph can be observed in Figure 30.

As a result, it was concluded that increasing validation dataset size have not aided our model in terms of training accuracy. Hence; in order not to sacrifice testing accuracy, the validation accuracy problem was accepted as is as we cannot increase the size of PAML dataset.

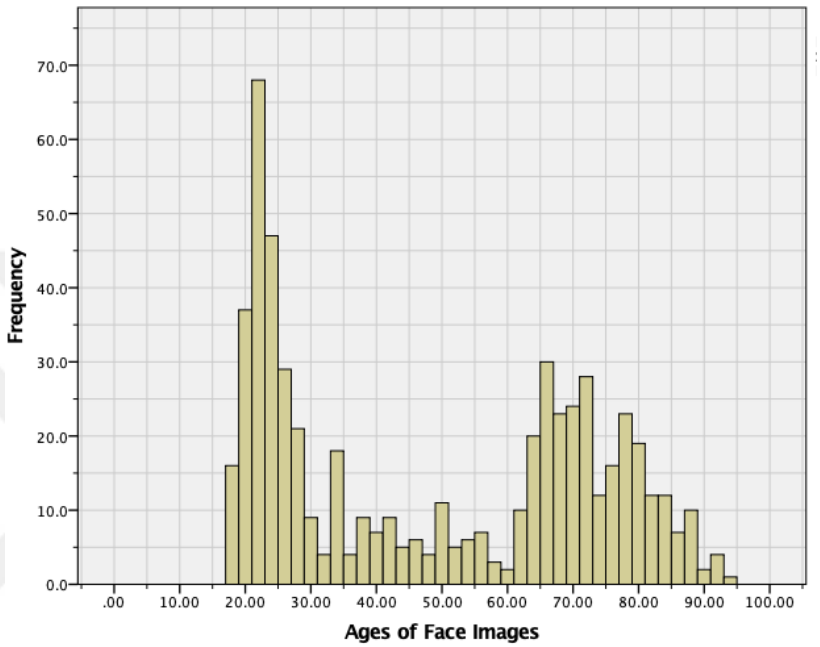

Figure 28: Distribution of face ages of images

Figure 29: Training accuracy percentages of the constructed ANN model. Red line represents training accuracy and blue line represents validation accuracy.

Figure 30: Training accuracy percentages of the constructed ANN model where validation dataset size was increased. Red line represents training accuracy and blue line represents validation accuracy.

In testing phase, the model has given satisfying reaction to the full 580 face images from PAML-AD. In Table 11 one might observe the performance metrics of ANN model.

Table 11: Performance metrics of ANN model.

ANN Performance Metrics	**Exact Match**	**Match With ±5 Age**	**Match With ±10 Age**	**Absolute Mean Prediction Error**	**Absolute Prediction Error Standard Deviation**
	63,45%	77,41%	85,34%	4,62	10,05

When the performance of the model is considered with respect to the age groups of the face images, following table appears:

Table 12: Performance metrics of the model, broken down to the age groups.

Age Prediction Ranges vs. Face Ages of Images		**Age Groups of Face Images**		
		Young Adult	**Adult**	**Elder**
Ranges	**Exact Match**	67,84%	53,00%	63,64%
	Match with ±5 Age	89,43%	59,00%	73,91%
	Match with ±10 Age	94,71%	66,00%	84,58%

When the validation problem and having unequal and high number of classification groups were to be considered, model's exact age prediction match is satisfyingly high. Especially when the sensitivity of the age prediction match was decreased, i.e. if the prediction is in ±5 or ±10 years range, the performance of model increases to 85%.

A candidate solution to the validation problem was to adjust PAML dataset in a way that it would have larger age groups with equal number of face images rather than individual age groups with unequal number of face images. In sections 4.2.2 - 4.2.5 it was attempted to fix the problem by adjusting the dataset.

4.2.2 Trained with using 10-adjusted groups

When the model was trained by adjusting PAML-AD in a homogenous manner, following results in

Table 14 were obtained. A total of 170 face images were used to train the model, i.e. 17 face images for each age group. Age groups are classified as follows:

Table 13: 10-adjusted age groups age group description table

Age Group Number	**Age Range**
1	18-26
2	26-33
3	33-40
4	40-48
5	48-56
6	56-63
7	63-70
8	70-78
9	78-86
10	86-100

ANN prediction results were as following:

Table 14: ANN results of 10-adjusted age groups

Neural Network Prediction Results	
Percentage of exact group prediction	31,03%
Percentage of predicted age's group is in +/- 5 years range with actual age of image	47,76%
Percentage of predicted age's group is in +/- 10 years range with actual age of image	61,72%
Absolute mean prediction error with respect to age group numbers	1,72
Standard deviation in absolute prediction error with respect to age group numbers	1,77

4.2.3 Trained with using 10-adjusted groups, only selected images tested

The results have changed as following, when only the images that were selected for training was tested for prediction:

Table 15: ANN results of 10-adjusted age groups when only trained faces were tested

Neural Network Prediction Results	
Percentage of exact group prediction	63,53%
Percentage of predicted age's group is in +/- 5 years range with actual age of image	77,06%
Percentage of predicted age's group is in +/- 10 years range with actual age of image	83,53%
Absolute mean prediction error with respect to age group numbers	0,75
Standard deviation in absolute prediction error with respect to age group numbers	1,23

4.2.4 Trained with using 5-adjusted groups

For 5-adjusted age groups, following age range table was used:

Table 16: 5-adjusted age groups age group description table

Age Group Number	**Age Range**
1	18-33
2	33-48
3	48-63
4	63-78
5	78-100

A total of 230 face images were used to train the model, i.e. 46 face images for each age group. ANN prediction results were as following:

Table 17: ANN results of 5-adjusted age groups

Neural Network Prediction Results	
Percentage of exact group prediction	55,86%
Percentage of predicted age's group is in +/- 5 years range with actual age of image	64,31%
Percentage of predicted age's group is in +/- 10 years range with actual age of image	74,14%
Absolute mean prediction error with respect to age group numbers	0,66
Standard deviation in absolute prediction error with respect to age group numbers	0,85

4.2.5 Trained with using 5-adjusted groups, only selected images tested

The results have changed as following, when only the images that were selected for training was tested for prediction:

Table 18: ANN results of 5-adjusted age groups when only trained faces were tested

Neural Network Prediction Results	
Percentage of exact group prediction	76,09%
Percentage of predicted age's group is in +/- 5 years range with actual age of image	84,35%

Percentage of predicted age's group is in +/- 10 years range with actual age of image	90,87%
Absolute mean prediction error with respect to age group numbers	0,35
Standard deviation in absolute prediction error with respect to age group numbers	0,68

4.2.6 Trained with PAML-AD, tested with FGNET-AD

Model's classification capabilities were tested on a dataset that the model have not encountered before. Prediction results of ANN model are as follows:

Table 19: ANN performance metrics for FGNET dataset.

ANN Performance Metrics	**Exact Match**	**Match With ±5 Age**	**Match With ±10 Age**	**Absolute Mean Prediction Error**	**Absolute Prediction Error Standard Deviation**
	0,70%	7,98%	14,77%	38,47	21,32

4.2.7 Trained with PAML-AD's neutral faces, tested with PAML-AD's happy faces

Since FGNET dataset's images were not as good as PAML-AD's images, happy faces were tested as different facial expression can create a uniqueness among training and testing images. Prediction results of ANN model are as follows:

Table 20: ANN performance metrics for PAML dataset's happy faces sub dataset.

ANN Performance Metrics	**Exact Match**	**Match With ±5 Age**	**Match With ±10 Age**	**Absolute Mean Prediction Error**	**Absolute Prediction Error Standard Deviation**
	29,84%	55,81%	67,83%	10,79	14,98

4.3 Behavioural Experiment – ANN Experiment Comparison of Results

The deep net trained over the same database exhibited the error distribution shown in the histogram below. As compared to human raters, the distribution showed higher kurtosis around the value zero, due to the larger number of cases where the model perfectly predicted the age of the person in the image. However, there were several outlier cases where the model performed extremely poorly.

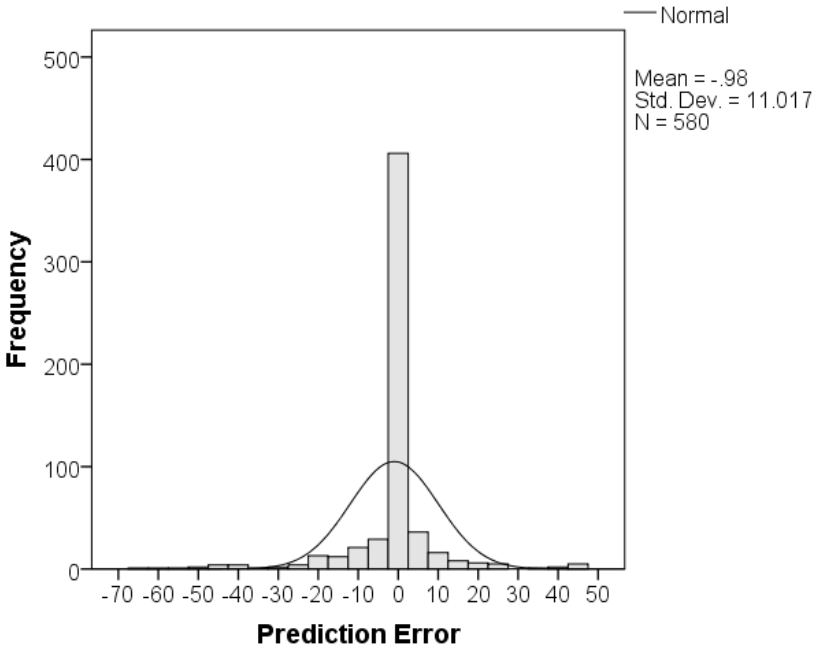

Figure 31: Error distribution histogram of deep net predictions.

Since multiple human ratings were obtained per image, and only a single prediction could be computed from the model, the human ratings were aggregated per image so as to make comparisons among human raters and the deep net model. For this purpose, we employed three different aggregations where the first one captures the best estimate made by a human rater, whereas the second and the third views captured the median and the mean estimates of the human raters for the corresponding image.

The boxplots show the distribution of estimation errors for the deep net model, best human estimation, median and mean of human estimations. The model includes several zero error cases, which led to an interquartile range of zero. Slightly larger interquartile ranges can be observed for the aggregated human estimations, where the best human estimation case is the closest to the model's interquartile range. However, there are also several outlier cases in the model distribution, where the model performed very poorly. Such extreme outliers are absent from human data.

Due to the presence of outliers and non-uniform error distributions, non-parametric Wilcoxon signed ranks tests were used to compare the model's and the human raters' accuracy. The test showed that the median error for the model (Median = 0, IQR=0) is significantly lower than the best human estimation (Median = 0, IQR=0), $z=-2.56$, $p<.01$, median human estimation (Median=.5, IQR=8), $z=-2.38$, $p<.05$, and mean human estimation (Median=.78, IQR=7.92), $z=-2.72$, $p<.01$ respectively.

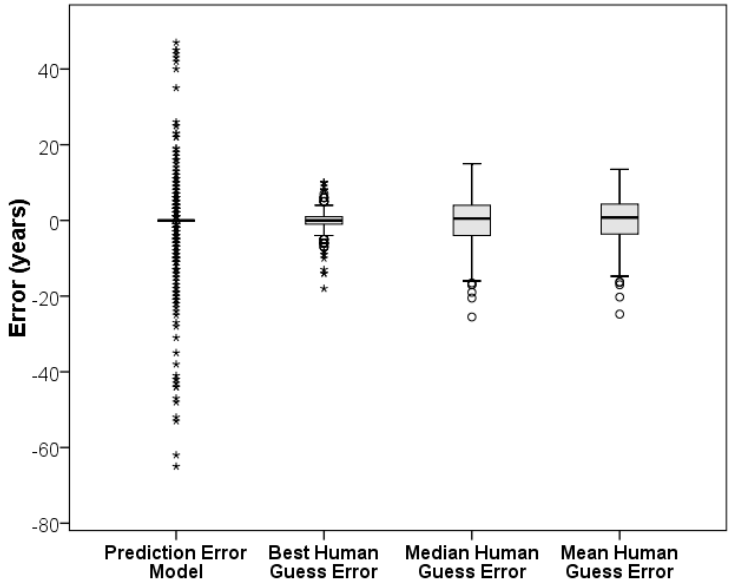

Figure 32: Boxplots of distribution of estimation errors for the deep net model, best human estimation, median and mean of human estimations.

We further investigated the relationship between the model and the human raters via correlation and regression analysis. The scatterplot below compares the age predictions obtained from the model and the real age of the people in the images. Although the majority of the predictions were on target, there are also significant deviations from the real age. In particular, there are cases where the real age is more than 60 but the model predicted 20 to 30, as well as cases where the real age is 60+ but the model predicted 20s. A significant and positive correlation was observed between the real age and the model's prediction, r=.89, p<.001.

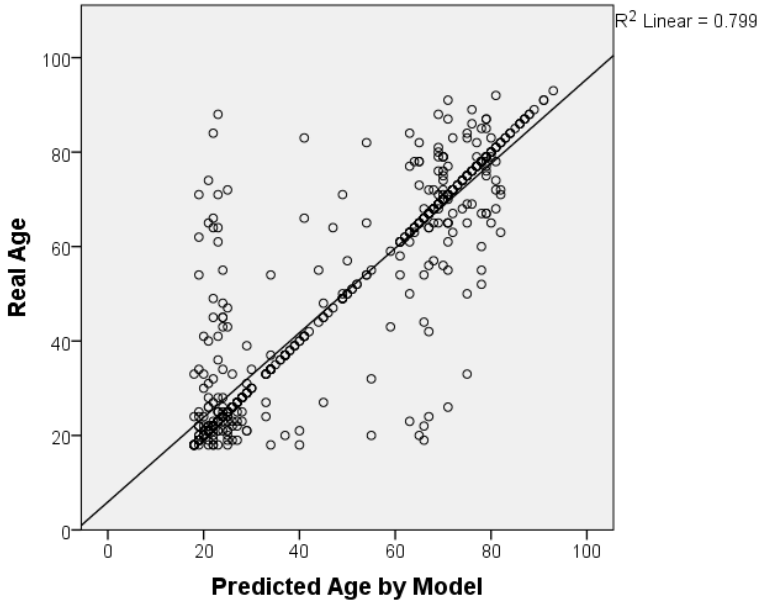

Figure 33: Scatterplot displaying age predictions obtained from the model and the real age of the people in the images.

The scatterplots below compare the real age with estimations obtained by the best human guessers, the median and the mean estimations. These aggregations can be considered to reflect a kind of wisdom of the crowd summaries of the data set. Significant positive correlations were observed for the best ($r=.99$, $p<.001$) median ($r=.97$, $p<.001$) and mean human ($r=.97$, $p<.001$) estimations. Although the number of perfect matches were lower, there were also no radically deviant predictions in the human ratings.

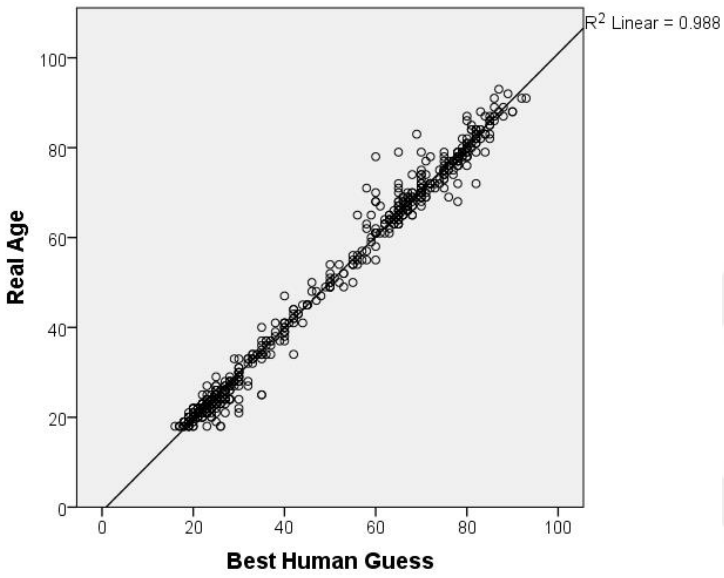

Figure 34: Scatterplot comparing real age of the face image shown and best human guess regarding the image.

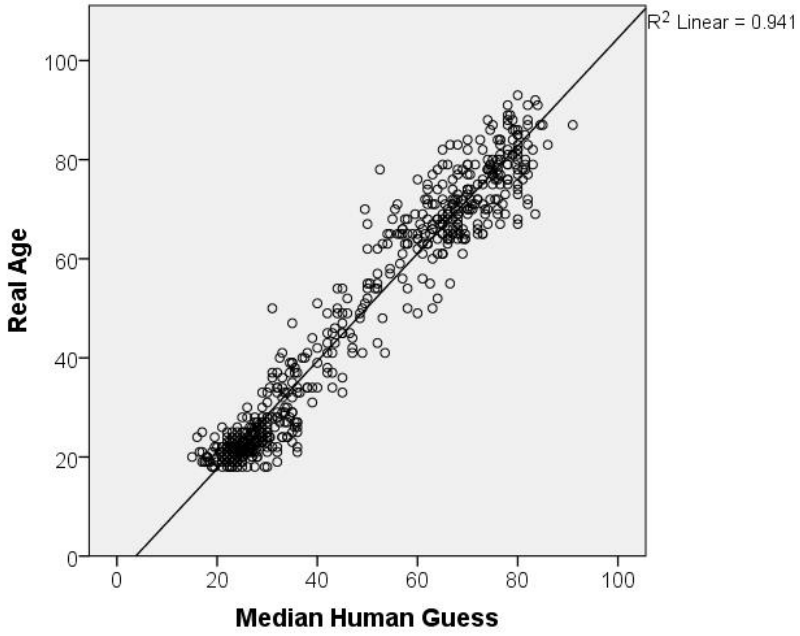

Figure 35: Scatterplot comparing real age of the face image shown and median human guess regarding the image.

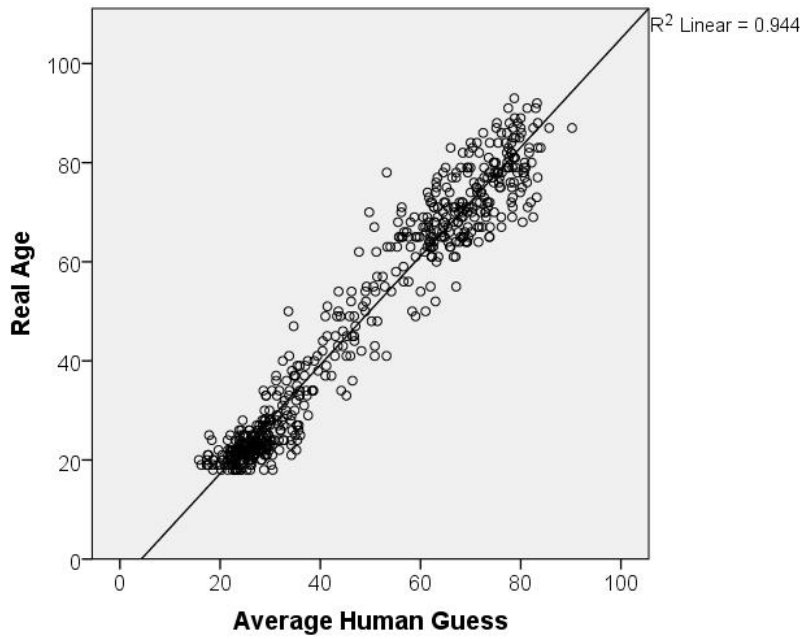

Figure 36: Scatterplot comparing real age of the face image shown and average human guess regarding the image.

We finally compared the model's predictions with human raters' estimations. Scatterplots suggest that there is considerable shared variability as indicated by positive and large Pearson correlations (i.e. $r=.89$, $p<.01$ for both the best and the mean human estimations). However, the model seemed to be oversensitive to the underlying age distribution of the pictures in our sample.

4.4 SVM Experiment Results

The results of SVM experiments will be reported in three forms; first Scikit's classification report, then confusion matrix and finally receiver operating characteristics.

For each experimental setting; which are original experiment, 5-adjusted groups and 10-adjusted groups, all three resulting reports will be given in their respective headings.

4.4.1 All PAML-AD images used as dataset

When image dataset is split in 80%-20% training-testing ratio, we had 464 images for training and 116 images for test dataset. In the following table the classification report that is generated by Scikit's metrics function can be observed.

Table 21: Classification report of SVM based classifier for whole dataset case.

labels	precision	recall	f1-score	support
18	1,00	1,00	1,00	5
19	1,00	0,50	0,67	2
20	1,00	1,00	1,00	1
21	0,90	0,75	0,82	12
22	0,88	1,00	0,93	7
23	0,80	0,80	0,80	5
24	1,00	0,67	0,80	6
25	1,00	1,00	1,00	3
26	0,00	0,00	0,00	2
27	1,00	1,00	1,00	3
28	1,00	1,00	1,00	2
29	0,50	1,00	0,67	1
31	1,00	1,00	1,00	1
33	1,00	1,00	1,00	3
34	1,00	0,50	0,67	2
37	1,00	1,00	1,00	2
38	1,00	1,00	1,00	1
39	1,00	1,00	1,00	1
40	1,00	1,00	1,00	1
41	1,00	1,00	1,00	3
42	1,00	1,00	1,00	1
45	1,00	1,00	1,00	2
47	1,00	1,00	1,00	1
48	1,00	1,00	1,00	1
49	1,00	1,00	1,00	1
50	1,00	1,00	1,00	1
54	1,00	1,00	1,00	1
55	1,00	1,00	1,00	1
61	0,00	0,00	0,00	1
63	0,67	1,00	0,80	2
64	0,50	1,00	0,67	1
65	1,00	1,00	1,00	1
68	1,00	1,00	1,00	1
69	1,00	1,00	1,00	4
70	0,50	1,00	0,67	1
71	0,80	1,00	0,89	4
72	0,33	1,00	0,50	2
73	1,00	1,00	1,00	1
75	1,00	1,00	1,00	1
76	1,00	1,00	1,00	1
77	0,00	0,00	0,00	1
78	1,00	1,00	1,00	2
79	1,00	1,00	1,00	5
80	1,00	1,00	1,00	1
82	1,00	0,75	0,86	4
83	1,00	1,00	1,00	2
84	1,00	1,00	1,00	2
85	0,00	0,00	0,00	0
87	1,00	0,67	0,80	3
88	1,00	1,00	1,00	1

Table 21 continued

91	1,00	1,00	1,00	2
92	1,00	1,00	1,00	1
accuracy			0,88	116
macro avg	0,86	0,88	0,86	116
weighted avg	0,90	0,88	0,88	116

In Table 21 precision column gives us the accuracy of positive predictions for each label, in our case labels are exact age points. Recall column is the fraction of positives that were correctly identified. F1-score column is a metric for comparing classifiers, it takes the harmonic mean of precision and recall columns.

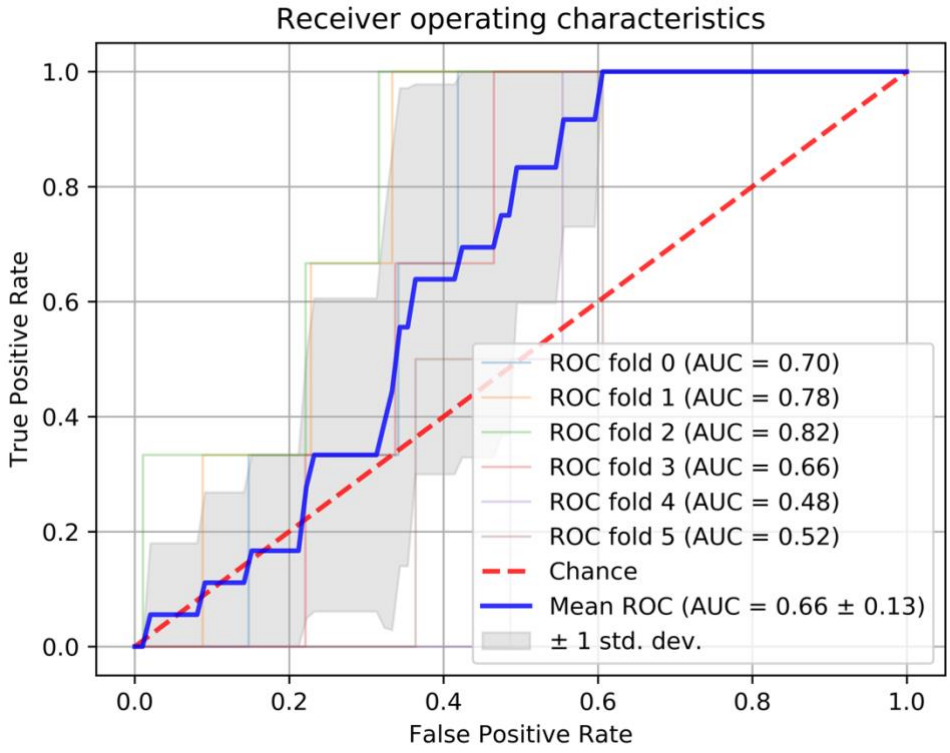

Figure 37: Receiver operating characteristics of whole dataset case.

In order to report a six-fold ROC, the dataset is randomly shuffled and selected (in our case k=6 times) and the false-positive and true-positive rates are calculated and drawn in Figure 37. By observing the random chance (red dashed line), standard deviation (grey shade), six ROC folds and mean ROC the performance of the classifier can be deduced to some extent.

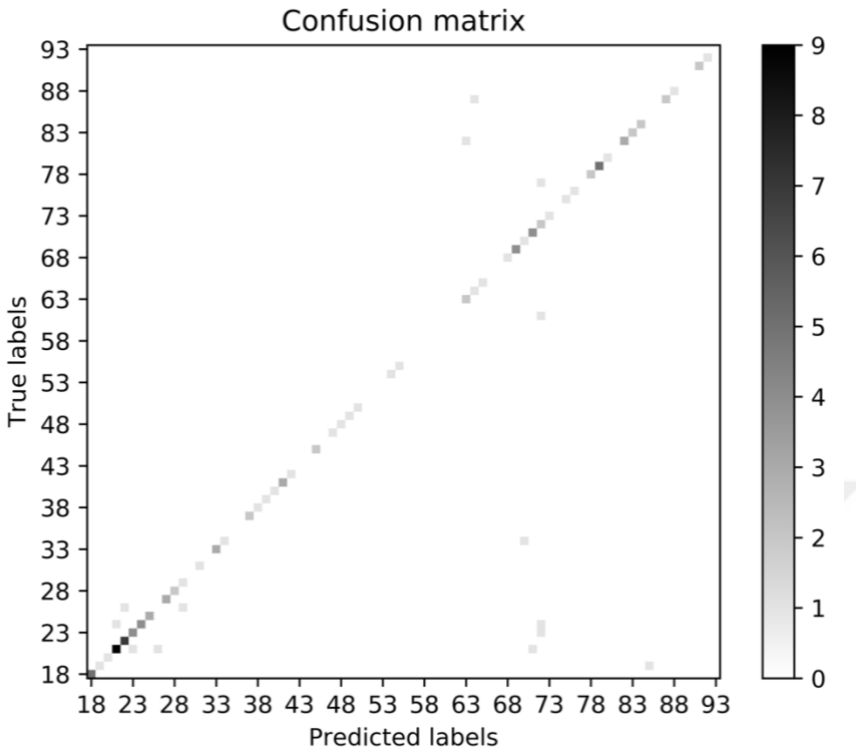

Figure 38: Confusion matrix of whole dataset case.

Confusion matrix is given for whole dataset case in Figure 38, yet as we have 75 cases the confusion matrix is not really clear for performance related conclusions.

4.4.2 10-adjusted groups used as dataset

When image dataset is split in 80%-20% training-testing ratio for 10-adjusted groups case, we had 136 images for training and 34 images for test dataset.

Table 22: Classification report of SVM based classifier for 10-adjusted groups case.

labels	precision	recall	f1-score	support
1	1,00	1,00	1,00	4
2	1,00	1,00	1,00	2
3	1,00	1,00	1,00	4
4	1,00	1,00	1,00	4
5	1,00	1,00	1,00	2
6	1,00	1,00	1,00	4
7	1,00	0,50	0,67	2
8	0,75	1,00	0,86	3
9	1,00	1,00	1,00	3
10	1,00	1,00	1,00	6
accuracy			0,97	34
macro avg	0,97	0,95	0,95	34
weighted avg	0,98	0,97	0,97	34

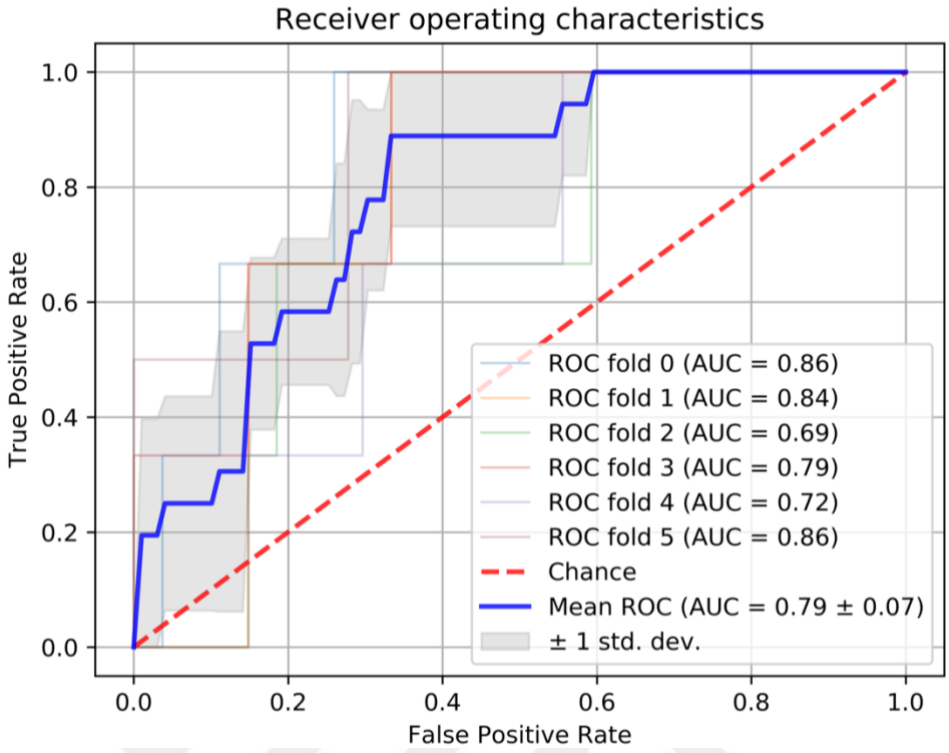

Figure 39: Receiver operating characteristics of 10-adjusted images case.

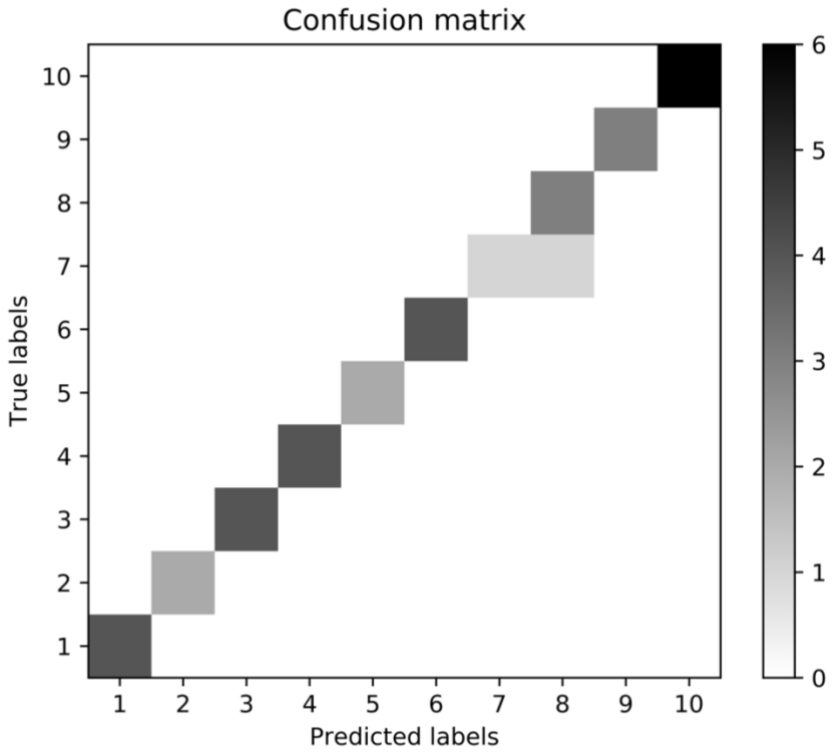

Figure 40: Confusion matrix of 10-adjusted images case.

4.4.3 5-adjusted groups used as dataset

When image dataset is split in 80%-20% training-testing ratio for 5-adjusted groups case, we had 184 images for training and 46 images for test dataset.

Table 23: Classification report of SVM based classifier for 5-adjusted groups case

labels	precision	recall	f1-score	support
1	0,89	1,00	0,94	8
2	1,00	1,00	1,00	5
3	1,00	0,92	0,96	13
4	1,00	0,90	0,95	10
5	0,91	1,00	0,95	10
accuracy			0,96	46
macro avg	0,96	0,96	0,96	46
weighted avg	0,96	0,96	0,96	46

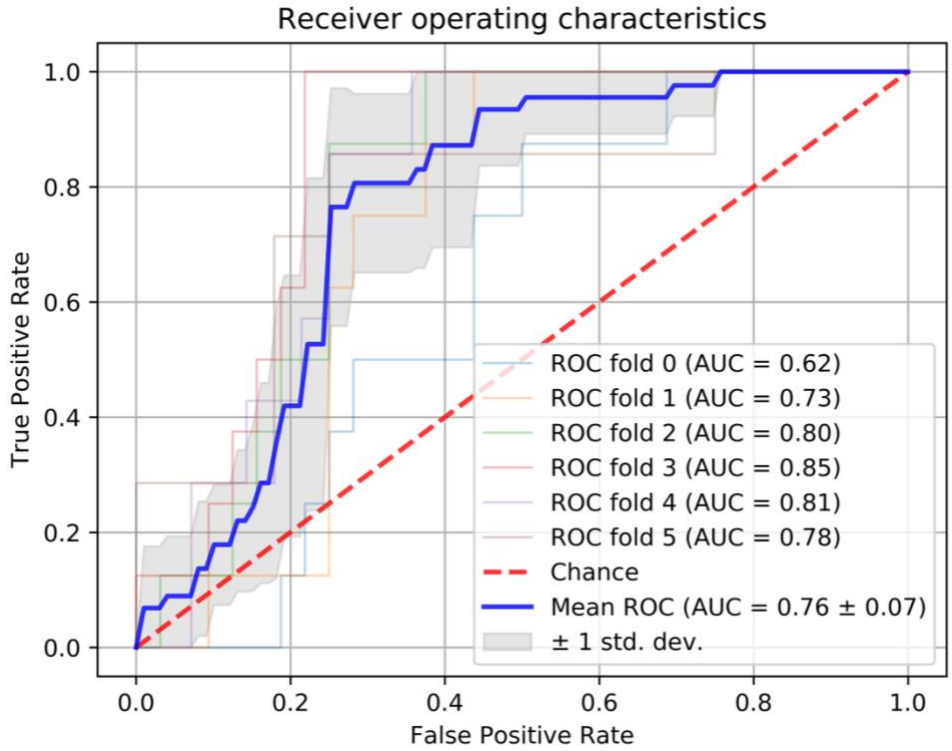

Figure 41: Receiver operating characteristics of 5-adjusted images case.

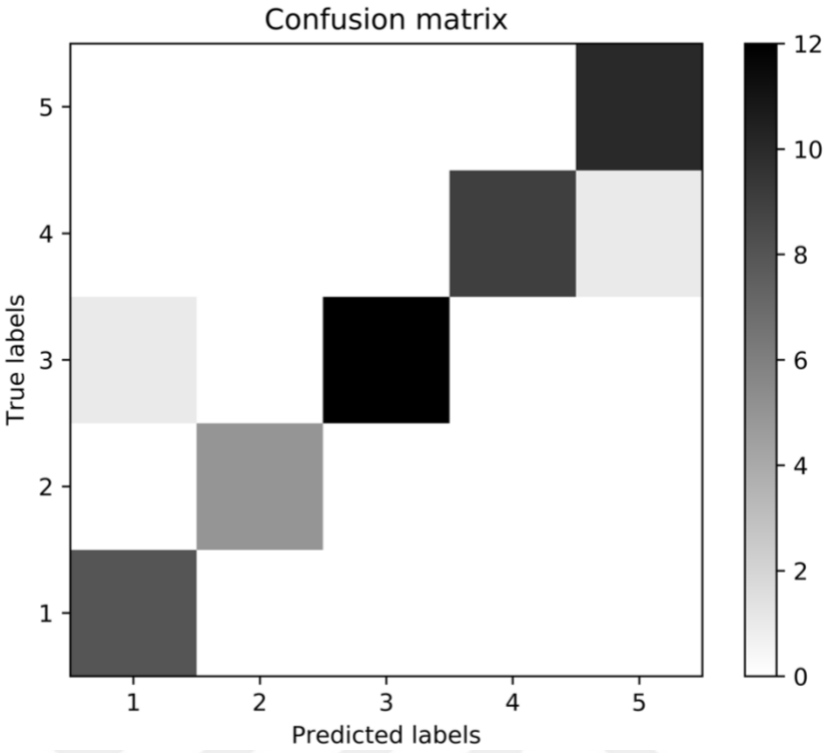

Figure 42: Confusion matrix of 10-adjusted images case.

CHAPTER 5

DISCUSSION & CONCLUSION

5.1 Behavioural Experiment

To compare our experiment's results and the results in literature firstly social-cognitive theories of the CAE should be discussed. Aforementioned Hugenberg et al.'s research had perspectives in in-group and out-group facial recognition, and they have suggested a CIM to acknowledge this in/out group recognition. Shortly Hugenberg et al. suggests that there exists a superior performance in recognizing in-group faces when compared to out-group faces. CIM in its core proposes that for in-group faces the perceiver (or rater) engages with an individuation process which is the product of categorization. In opposition to in-group faces, out-group faces quickly categorized as an alien, i.e. out-group, face. Henceforth, Hugenberg et al.'s CIM may provide a viable explanation to the existence of CAE (Hugenberg et al., 2010).

In our experiment, we have observed a positive relationship between measured error (guessed age – actual age) and objective distance (self-age – actual age) and in contrast to this almost zero correlation was observed between measured error and perceived distance (self-age – guessed age). Yet, a strong positive correlation between objective and perceived error was observed, and this suggested that participant's age recognition (or categorization) were consistent with respect to the perceived age of the image they have seen. Since these results neither supported nor hindered our initial hypothesis further tests were devised. By splitting data into 6 groups (details were given in Section 4.1: Behavioural Experiment Results), the in-group and out-group effects were tested. Results of the tests have validated the findings in the literature; age estimations were better for the ages that is closer to the guesser's own age, and this finding is the basis of the CAE hypothesis that is investigated.

More on individuals' age guessing strategies; in the literature when recognizing facial age, age group levels, such as *children, young adult, middle-aged adults, adults, older adults, elder,* were utilized frequently. In our experiment, instead of using vague age groups fixed age points were used. When compared to the related literature where age groups are typically utilized, the participants' decision process was expected to be harder as it was required for them to match the face they have seen to a certain age point. Hence; mostly influenced by Hugenberg et al.'s CIM, we have suggested that when humans are guessing ages, regardless of the bin size, their strategy is to group faces in order to categorize them. Then, they would tend to individualize them according to their in-group and out-group facial age memory capabilities.

5.2 ANN Based Experiments

First major obstacle that have appeared during the construction of an ANN experiment was the size constraint of the facial aging dataset that was used. As there weren't many open source datasets that is directly designed for facial aging purposes; we opted for

the PAML-AD. PAML-AD would be a perfect dataset for especially behavioural experiments, as we have used it without any major problems. But for experiments involving ANNs, a huge number of data points is a must, which turned out to be a limiting factor given the fact that the size of the PAML-AD was just 580 images.

As mentioned in earlier sections; even though the training accuracy of the model was high (above 90%), the validation accuracy was significantly low (below 20%). This issue was the implication that our model memorizes each face image instant, and any other facial features that it has learned, instead of actually learning the difference between each classification, in our case age points.

Increasing the validation dataset size and changing the learning parameters in order to deal with this problem did not help us as the validation and training accuracies became unstable after changing parameters and the dataset size. For comparison of the accuracies before and after the supposed adjustments, please refer to Figure 29 and Figure 30.

As our classification groups are exact ages of the face images, the frequency of the classification groups was not uniformly distributed. A histogram of the classification groups was provided in Figure 28. In order to eradicate this non-uniform distribution, experiments in sections 4.2.2, 4.2.3, 4.2.4 and 4.2.5 were devised. Basically, in that four experiments, the face images were grouped with respect to the number of images and the total distribution was attempted to be normalized as much as possible.

In the table below, a performance-based comparison of those four experiments and the original experiment is given. Moreover, in Figure 43 the performance comparison is visualized in the form of bar graphs.

Table 24: Performance-wise comparison of ANN experiments.

Experiment name	**Original experiment**	**10-adjusted age groups**	**10-adjusted age groups selected images**	**5-adjusted age groups**	**5-adjusted age groups selected images**
Percentage of exact group prediction	63,45%	31,03%	63,53%	55,86%	76,09%
Percentage of predicted age's group is in +/- 5 years range with actual age of image	77,41%	47,76%	77,06%	64,31%	84,35%
Percentage of predicted age's group is in +/- 10 years range with actual age of image	85,34%	61,72%	83,53%	74,14%	90,87%
Absolute mean prediction error with respect to age group numbers	4,62	1,72	0,75	0,66	0,35
Standard deviation in absolute prediction error with respect to age group numbers	10,05	1,77	1,23	0,85	0,68

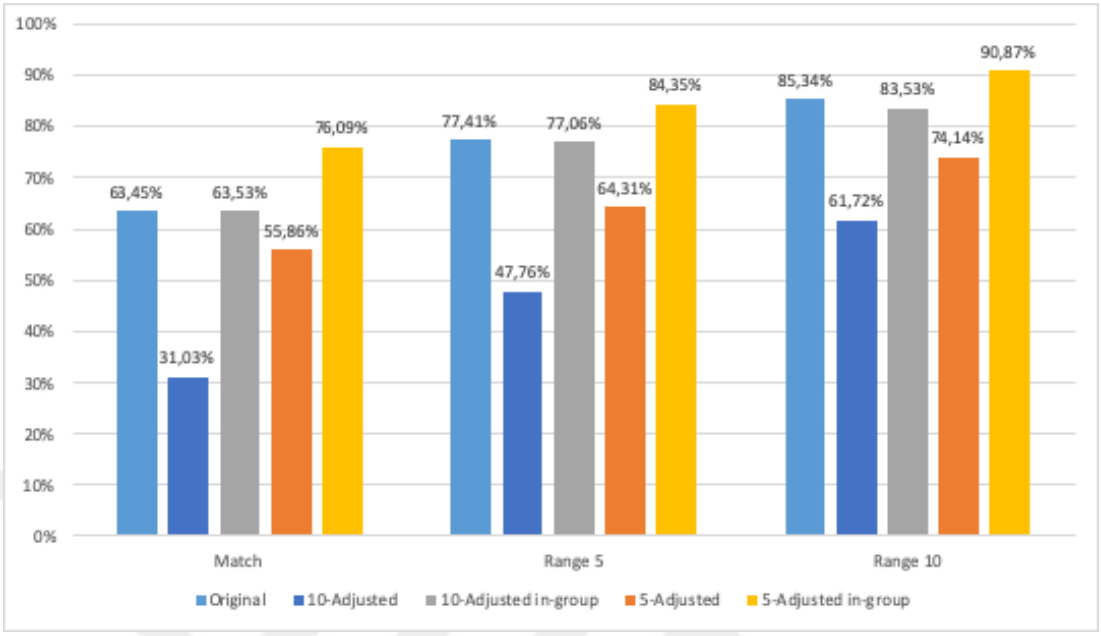

Figure 43: Performance-wise comparison graph of ANN experiments regarding different group sizes.

As seen both from Table 24 and Figure 43 adjusting the group size did not yield significant accuracy improvements. Only in exact match accuracy, there was a noticeable difference among original experiment and 5-adjusted in-group experiment which used 46 images for five age groups. In the light of this result; it can be speculated that we would need at least 46 images for each age group, of course 46 is an arbitrary and speculative number of images. So, if a full-precision age prediction experiment based on ANNs would be considered; our latest findings imply that we need at least 3450 images, i.e. age range between 18-93 and 46 images for each age point, just for training purposes.

To check the validity of our model's age rating ability, experiments in sections 3.2.4.6 and 3.2.4.7 were devised. The premise of these two experiments were to train our model with PAML-AD and then test it with completely new images; one with happy emotion faces in PAML-AD and other with FGNET-AD.

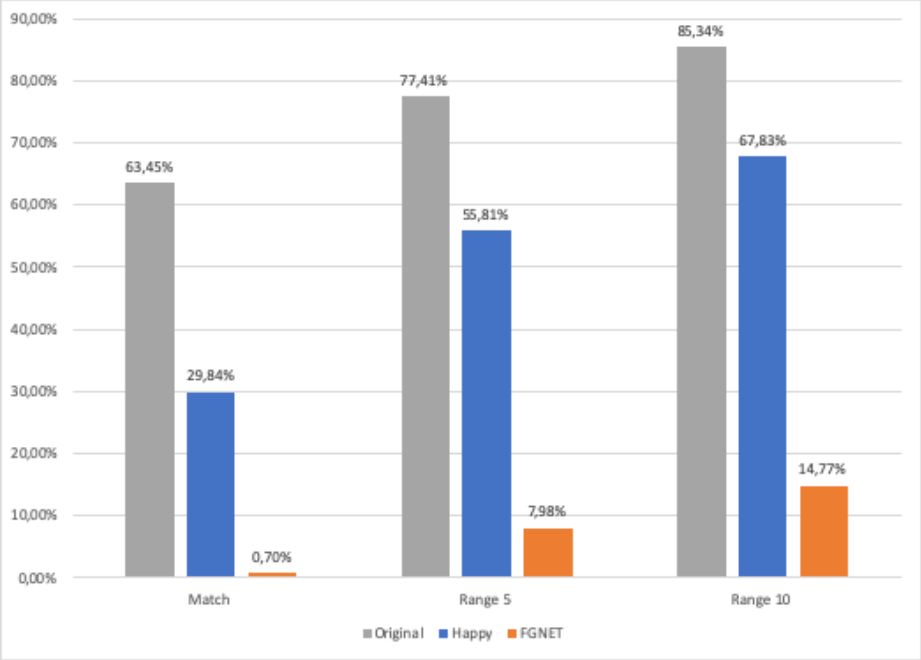

Figure 44: Performance-wise comparison graph of experiments regarding unique image prediction datasets.

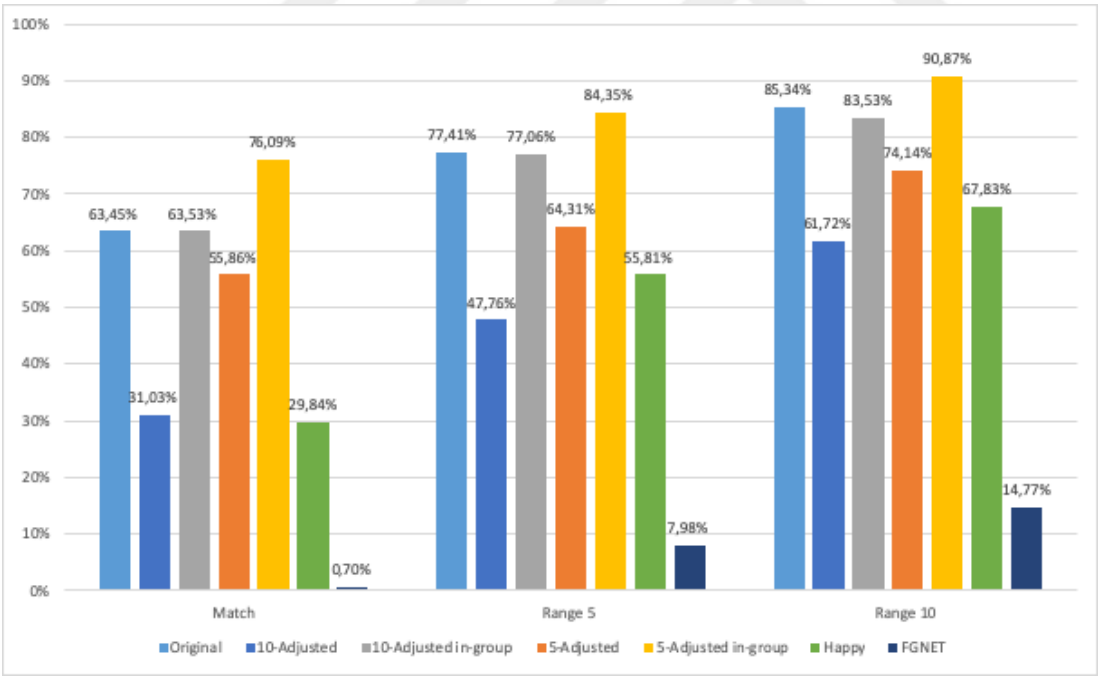

Figure 45: Performance-wise comparison graph of all experiments discussed above.

As seen from Figure 44, the model fails when making predictions on FGNET-AD but it's performance on PAML-AD's happy faces is noticeably well. When comparing all experiments' performances, that is given in Figure 45, it is somewhat clear that the performance of the model is heavily dependent on quality of the images and quantity of the images. FGNET experiment fails, because the quality of the images is mostly low. When acceptable quality images were to be fed into the model, in the shape of

happy faces in PAML-AD, the performance increases and even exceeds 10-adjusted groups experiment.

As deep-nets have a “black-box” structure by their nature, we have checked layer outputs of our model. Our aim was to identify any similarities between how humans perceive facial features and how our model was trained to perceive such features, if it did train itself. In following figures random images were fed from our dataset to the model and our model’s deep layer outputs were saved as separate files. Please refer to Figure 16 for our model’s architecture.

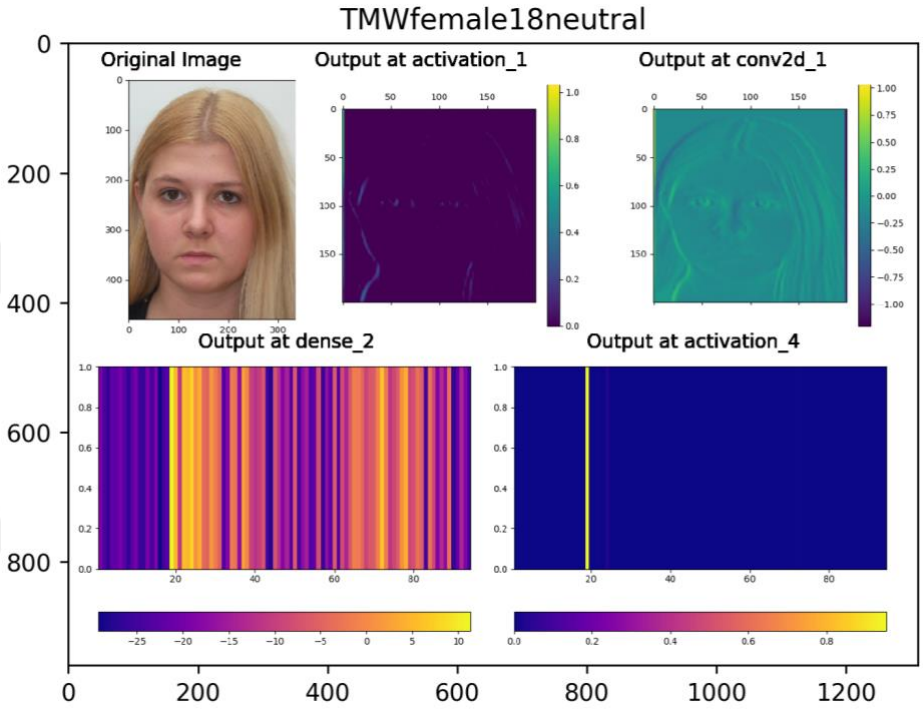

Figure 46: Deep-net layer outputs of an 18-year-old female facial image. Output images are titled with respect to their layers.

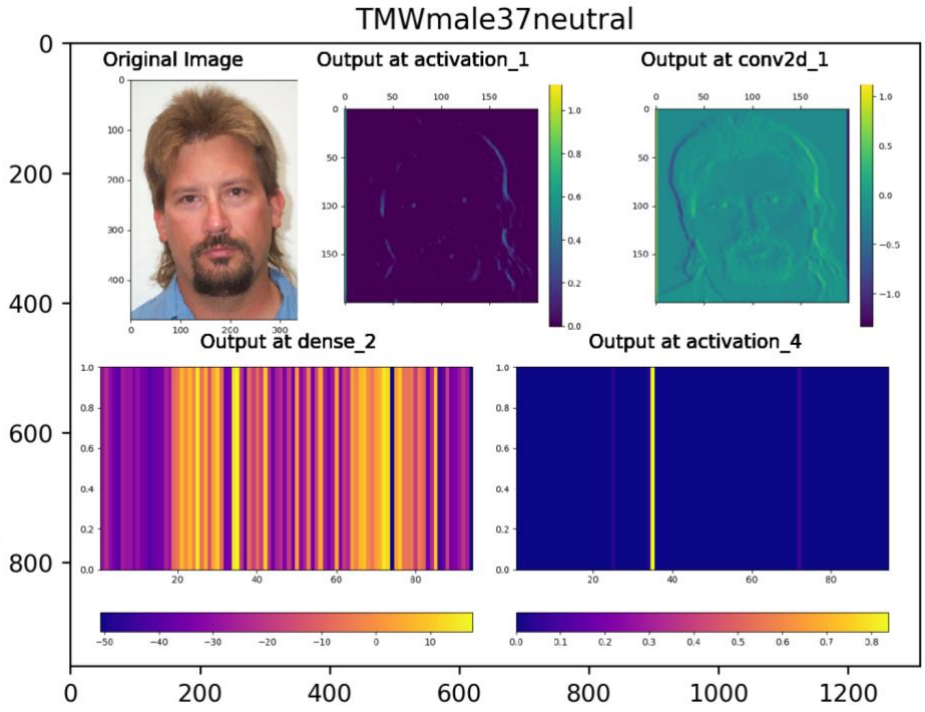

Figure 47: Deep-net layer outputs of a 37-year-old male facial image. Output images are titled with respect to their layers.

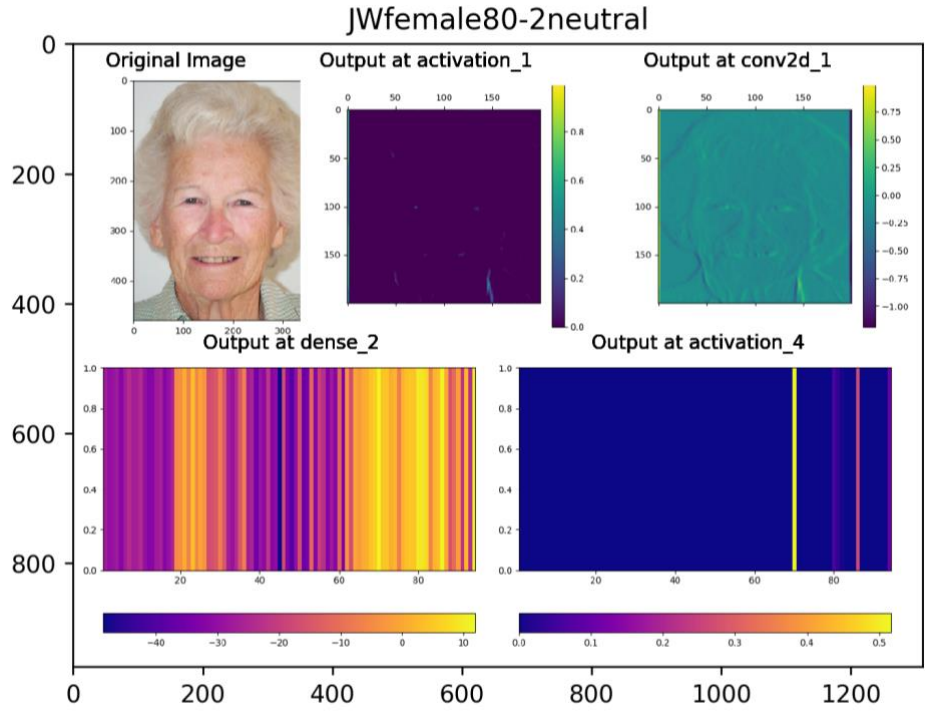

Figure 48: Deep-net layer outputs of an 80-year-old female facial image. Output images are titled with respect to their layers.

When Figure 46, Figure 47 and Figure 48 observed closely; our model attempts to detect facial features, as they start to appear on "activation_1" layer. Since

aforementioned layer is one of the shallowest layers, it can be deduced that the model's strategy of age prediction involves linking facial features with age bins (in our case each age point is also an age bin). If the model's strategy were compared with previous research on facial features, Ekman and Friesen's Facial Action Coding System (FACS) is one of the most substantial facial feature recognition systems that was proposed (Ekman & Friesen, 1978). In addition to Ekman and Friesen, Pantic and Rothkrantz had proposed an approach to automated facial action recognition (Pantic & Rothkrantz, 2004). As both of the researches involve facial features, it was more than accurate to compare and contrast these two approaches and our model's strategy. In the following figure, Pantic and Rothkrantz's facial feature point map can be observed.

Moreover, as the images gathered from intermediate layer activation outputs are not really clear, Grad-CAM (Selvaraju et al., 2017) is utilized to visualize the intermediate layer activations. In the following figure, the Grad-CAM outputs of respective layers are printed on top of input images.

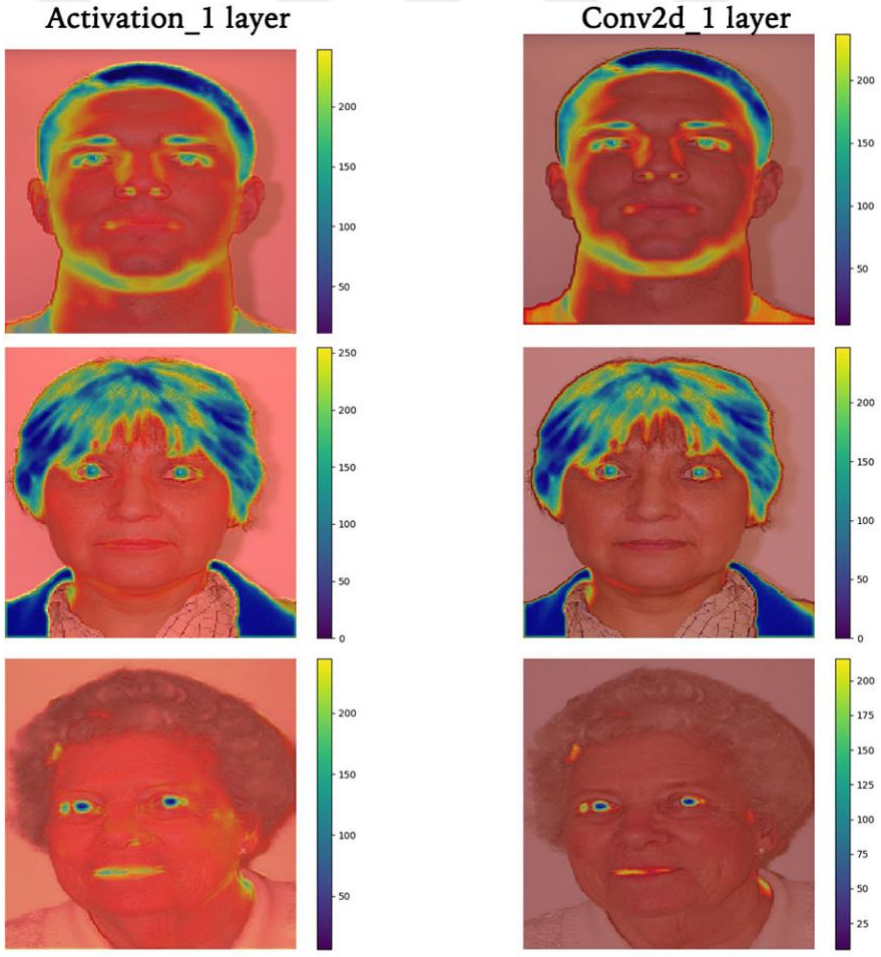

Figure 49: Grad-CAM outputs of selected images from first couple of 2D image matrix-based layers. Please notice that red tint is the no-activation baseline for Grad-CAM and activation colour bar is given for each image on the right of them.

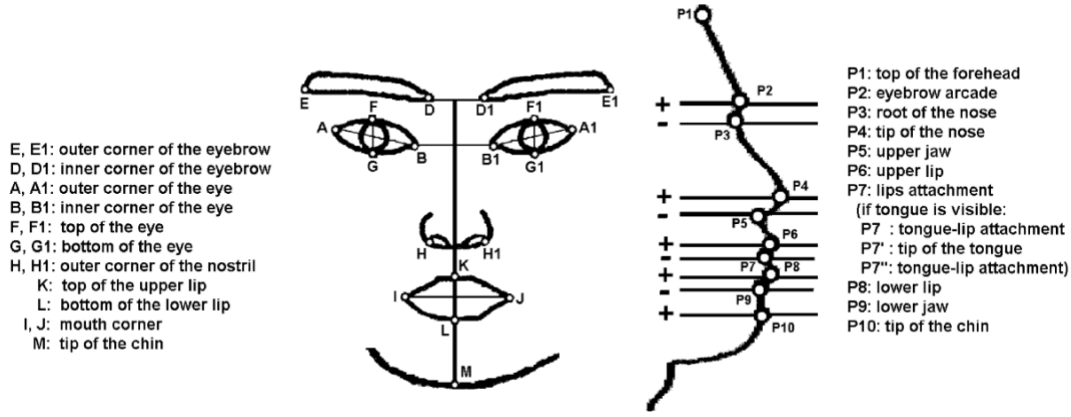

Figure 50: Pantic and Rothkrantz's facial feature point map with acknowledgements (Pantic & Rothkrantz, 2004, p.1452).

When Figure 50 was observed, the importance of eyes and nostrils had similarities with our model's approach as they both used those features in a common basis; the detection of such features can be observed in Figure 46, Figure 47 and Figure 48, furthermore in Figure 48 it can be observed that the model was somewhat confused and it have yielded an indecisive output as compared to Figure 46 and Figure 47.

The model had such indecisiveness for other facial images as well. When the indecisiveness is somewhat similar to Figure 48 it is understandable that the model is trying to elaborate the facial features to the age bins and the features do not possess obvious cues. Hence, the layer outputs of the model had activations all over a larger age bin.

Moreover, when the Grad-CAM outputs in Figure 49 are taken into attention, it is clear that our model is attempting to take utmost importance to eye, mouth and nose areas. Also, we have deduced that the model attempted to make classifications by using hair area for most of the input images presented to it. At the end, what we can argue regarding deep-layer Grad-CAM and activation output images is that even a naïve classifier attempts to classify human facial age by utilizing facial features; especially the ones that is reported in literature and used in non-machine learning classifiers. Hence by adapting an input image augmenting module, which will augment the ages in a way that important facial features will stand out more, achieving a ML-based facial age classifier (and also a CAE detector) is inevitable.

To sum up and propose further research notes on ANN experiment discussion, pre-processing the input data in order to enhance the human-likeness of the model via utilizing facial features that are aforementioned, and also can be seen in Figure 50, is one of the prominent suggestions that we can present. As our experiment had a naïve approach to the ANN model, we decided not to delve deep into pre-processing. Nevertheless, with exaggerating the facial features by formulas that would warp the facial images that would be used as inputs to the ANN model. Apart from facial features proposed by Pantic and Rothkrantz's facial feature point map, to propose an even better approach to the issue, the features mentioned in Section 1.3: Facial Aging should be utilized as well.

Lastly, we have attempted to cure the overfitting problem of our model. One of the first efforts was to feeding noise and images of completely unrelatable items into our model. Yet both methods failed as feeding noise yielded completely random images, and when unrelatable item's images were fed the model consistently associated their outputs to an unintelligible age classification group. Another attempt that we have hypothesized was to utilize a dataset that has more than enough facial images with age information encoded in it. But we have not had the chance to test this hypothesis, since the datasets that have a large number of facial images with age information is either non-existent or the researchers that own the datasets do not allow open-access to them.

5.3 SVM Experiment and Comparison with ANN

As the ANN was thought to be an over-powerful tool for a classification problem that looks rather trivial, SVM based experiments were designed. The design process was not rather hard, as all the image datasets for three cases (whole dataset, 5-age-group, 10-age-group) were ready to use as they have prepared for ANN. The results are reported in the forms of classification reports, ROCs and confusion matrices of three cases in SVM experiment in Section 4.4.

In whole dataset case just looking at classification report and concluding that the classification performance of SVM is enough for our purposes would be rather biased. Hence, first observing ROC in Figure 37 we can see that mean ROC is lying around random chance of classification line up to around 0.4 rate, but then it skews to the true positive rate. The area under the curve for mean ROC is 0,66 (with a std. deviation of 0,13), so actually our classifier can distinguish positive and negative cases with a 66% chance.

Moving on to CM in Figure 38, as we had labels that are representing ages 18-92 the CM is not really visible for this large label case. What we can deduce is in test dataset the classifier seems to classify with a high chance of having true positive cases.

For 5-adjusted and 10-adjusted groups, the classification reports show that the classifier is working with a rather higher performance. The mean ROC of 10-adjusted case yields an AUC of 0,79 which was one standard deviation higher than whole dataset case's AUC. So, the performance is significantly increased when the number of classification labels are reduced to 10 from 75. CM of 10-adjusted case was a little bit clearer and it shows a more stable true positive identification performance.

In 5-adjusted case, AUC of the mean ROC was 0,76; hence the performance of the classifier did not significantly change when the label number is reduced to 5 from 10. Yet again, CM showed a stable true positive identification performance.

When investigating SVM performance we have deduced that for large number of classification labels (such as our whole dataset case there were 75 classes) the performance is considerably low. In spite of that, for fewer classification labels (i.e. our 5-adjusted and 10-adjusted cases) SVM performance is significantly higher. Hence, when we compare ANN and SVM performances both of them had acceptable performances for fewer classification labels. But, for whole dataset case SVM's

performance is not convincing and satisfactory whereas ANN has a somewhat satisfactory performance for the case.

Hence, we have decided that for fewer labels using ANN can be over-kill. But for a large number of classification labels, using ANN would be a better choice performance-wise.

5.4 Contrasts Between Behavioural and ANN Based Experiments

Model's prediction pattern has both similarities and dissimilarities with how humans perceive facial age. From the very shallow nodes, the model attempts to pay close attention to facial features. Referring to our previous arguments about the existing link between facial features and facial age perception, our model's attempts on predicting faces by elaborating facial features can be considered as a close resemblance to the human facial age recognition strategies.

Regarding age prediction, there is an easily observable difference on human and ANN guessing strategies. As seen from the comparative boxplot in Figure 51, both human participants' and ANN model's mean prediction error was concentrated around 0; but for human participants the mean prediction error was in a larger interval when compared to ANN model. This might be the result of CIM occurring in participants which we have discussed in Section 5.1: Behavioural Experiment. In which during facial recognition process, in our case facial age recognition/prediction, the participants tend to first categorized faces into age groups even though they were not directed or motivated to do so. In the contrary, they were assigned to give exact age point guesses to the face images they have seen. Henceforth, we have asserted that individuals innately construct age groups in their facial (age) recognition mechanisms during cognition.

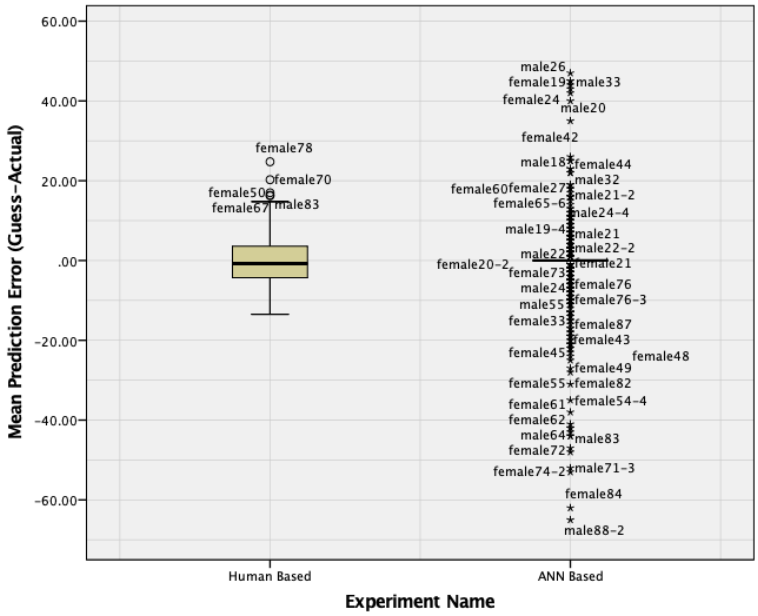

Figure 51: Boxplot comparison of mean prediction error of two experiment types.

To extend the discussion above on ANN experiment, just like individuals, our model was instructed to guess exact age points for the images that was presented. But unlike individuals, it attempted to give exact age points that can also be seen in Figure 51. When both of the experiments were to be compared in the scope of this "facial age guessing with using intervals" argument, in human based experiment this effect was clearly observed. Whereas in ANN based experiment, as expected, our model has given machine-like outputs to the facial image inputs.

5.5 Conclusion and Proposed Future Work

In Russell and Norvig's influential book on AI, the foundations of AI are investigated in an interdisciplinary manner. As they have reported early findings and aspirations generated from these early-stage advances that researchers had, the problems that have arisen have not omitted in the book. First difficulty that Russell and Norvig argue is that early AI programs typically contained little to no knowledge about the subject matter they have designed to work on. Second reported problem is that early attempts often covered the basic facts and the larger problems would be solved via scaling up the designed basic solver. And they have concluded that this scaling up problem is a matter of time and as the hardware would evolve, the problems will be solved. Yet, this foresight did not come true; as the third reported difficulty involved the fundamental limitations on baseline structures of generating intelligent behaviour. The perceptron created in early-stage research had the ability to learn and represent, but in an insignificant way (Russell & Norvig, 2010).

To sum up our research with using a perspective gained from Russell and Norvig, we have argued on some of the key points they have pointed out by adapting them on the research. Then, starting with weak/strong AI discussion is a nice point for our experiments' sake. We have easily claim that the ML based classifiers we have designed are in line with weak AI position, as by nature the ML based classifiers' aim is to distinguish between inputs that are given to them. Moreover, what we have intended besides classification performance was to identify what do the intermediate-layers do to classify age from face images. In this perception, neither side in weak/strong AI discussion clearly fit into our purposes (Russell & Norvig, 2010).

Yet; with a naïve classifier which have not had any idea about what is the concept of age or the concept of face, we have observed that with only 580 images that are not evenly distributed among an age range (i.e. having bias on some age points) it has managed to learn facial structures and tried to guess ages with utilizing those structures. Surprisingly (or actually unsurprisingly), this ploy is virtually similar to the one that humans use when guessing ages. In observing the activations of intermediate layers, we stand roughly satisfied with the results that we have achieved.

Regarding future work, a pre-processing via augmenting facial features were proposed in early sections. This proposal might increase the accuracy of the classifier, as the points that should be learned by the classifier in order to achieve its goal would be implicit when compared to the non-pre-processing cases. But for this proposal to work properly, constructing a solid facial feature extractor and augmenter designed upon the human based hypotheses reported in literature is key. More on future work, an eye-tracker experiment can be appended to the behavioural experiment we have

constructed in order to compare if the activations of intermediate layers of a naïve classifier will be in line with actual human data which are not really naïve. Also, as this comparison have already mentioned in our work with using experiments done in literature, the proposed eye-tracking study's results can be implemented as a pre-processing application via augmenting input images with respect to the first few saccades that are consistent among all participants that are recorded during facial age guessing.

Last but not least, other ML architectures are welcome to be tested in order to address the research questions that we have come up with. More on that, not only architectures but other methods that are trending in ML and DL, i.e. pruning, can be utilized in order to overcome the lack of having a satisfactorily large dataset.

REFERENCES

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., S. Kanwal, M., … Lacoste-Julien, S. (2017). *A Closer Look at Memorization in Deep Networks*.

Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*. https://doi.org/10.1037/0033-295X.94.2.115

Bindemann, M., Doherty, M. J., Burton, A. M., Schweinberger, S. R., & Langton, S. R. H. (2007). The control of attention to faces. *Journal of Vision*. https://doi.org/10.1167/7.10.15

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. https://doi.org/10.1145/130385.130401

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., … Xie, Z. (2018). Deep Learning and Its Applications in Biomedicine. *Genomics, Proteomics and Bioinformatics*. https://doi.org/10.1016/j.gpb.2017.07.003

Chiroro, P., & Valentine, T. (1995). An Investigation of the Contact Hypothesis of the Own-race Bias in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*. https://doi.org/10.1080/14640749508401421

Dakin, S. C., & Watt, R. J. (2009). Biological "bar codes" in human faces. *Journal of Vision*. https://doi.org/10.1167/9.4.2

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. *CA: Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From Appraisal to Emotion: Differences among Unpleasant Feelings. Motivation and Emotion*. https://doi.org/10.1007/s10751-008-9818-2

Frisby, J. P., & Stone, J. V. (2010). *Seeing: The Computational Approach to Biological Vision*. *MIT Press* (2nd Editio). Cambridge, MA. https://doi.org/10.1002/col.21817

Guo, G. (2013). Age prediction in face images. In M. Fairhurst (Ed.), *Age Factors in Biometric Processing* (pp. 231–251). UK: The Institution of Engineering and Technology. https://doi.org/10.1049/pbsp010e_ch13

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised Learning BT - The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.) (pp. 485–585). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7_14

Hershler, O., & Hochstein, S. (2005). At first sight: A high-level pop out effect for faces. *Vision Research*. https://doi.org/10.1016/j.visres.2004.12.021

Hole, G., & Bourne, V. (2010). *Face Processing: Psychological, Neuropsychological, and Applied Perspectives*. New York: Oxford University Press Inc., New York. Retrieved from https://books.google.com/books?id=7UicAQAAQBAJ&pgis=1

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The Categorization-Individuation Model: An Integrative Account of the Other-Race Recognition Deficit. *Psychological Review*. https://doi.org/10.1037/a0020463

Jakkula, V. (2006). Tutorial on Support Vector Machine (SVM). *School of EECS, Washington State University*. https://doi.org/10.11648/j.acm.s.2017060401.11

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*. https://doi.org/10.1146/annurev-vision-082114-035447

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*. https://doi.org/10.1038/nature14539

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. https://doi.org/10.1109/5.726791

Li, Y., & Tse, C. S. (2016). Interference among the processing of facial emotion, face race, and face gender. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2016.01700

Lin, S., Cai, L., Lin, X., & Ji, R. (2016). Masked face detection via a modified LeNet. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2016.08.056

Luck, S. J. (Steven J. (2005). *An introduction to the event-related potential technique*. *CEUR Workshop Proceedings*. https://doi.org/10.1118/1.4736938

MacLin, O. H., Van Sickler, B. R., MacLin, M. K., & Andrew, L. (2004). A Re-examination of the Cross-race Effect: The Role of Race, Inversion, and Basketball Trivia. *North American Journal of Psychology*.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/h0028434

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*. https://doi.org/10.1007/BF02478259

Meissner, C. A., & Brigham, J. C. (2001). Thirty Years of Investigating the Own-Race Bias in Memory for Faces: A Meta-Analytic Review. *Psychology, Public Policy, and Law*. https://doi.org/10.1037/1076-8971.7.1.3

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli.

Behavior Research Methods, Instruments, and Computers. https://doi.org/10.3758/BF03206543

ng, W. J., & Lindsay, R. C. L. (1994). Cross-race facial recognition: Failure of the Contact Hypothesis. *Journal of Cross-Cultural Psychology*. https://doi.org/10.1177/0022022194252004

Panis, G., Lanitis, A., Tsapatsoulis, N., & Cootes, T. F. (2015). Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*. https://doi.org/10.1049/iet-bmt.2014.0053

Pantic, M., & Rothkrantz, L. J. M. (2004). Facial Action Recognition for Facial Expression Analysis From Static Face Images. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *34*(3), 1449–1461. https://doi.org/10.1109/TSMCB.2004.825931

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., … Worek, W. (2005). Overview of the face recognition grand challenge. In *Proceedings - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*. https://doi.org/10.1109/CVPR.2005.268

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*. https://doi.org/10.1037/a0025750

Ricanek Jr, Karl; Mahalingam, G., & Albert, A. Midori; Bruegge, R. W. V. (2013). Human face ageing: a perspective analysis from anthropometry and biometrics. In M. Fairhurst (Ed.), *Age Factors in Biometric Processing* (pp. 93–116). The Institution of Engineering and Technology.

Rivolta, D. (2014). *Prosopagnosia: When All Faces Look the Same*. *Cognitive Systems Monographs*. https://doi.org/10.1007/978-3-642-40784-0_3

Robbins, R., & McKone, E. (2007). No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition*. https://doi.org/10.1016/j.cognition.2006.02.008

Rosebrock, A. (2016). LeNet – Convolutional Neural Network in Python. Retrieved August 10, 2018, from https://www.pyimagesearch.com/2016/08/01/lenet-convolutional-neural-network-in-python/

Rosebrock, A. (2017). Image classification with Keras and deep learning. Retrieved May 10, 2018, from https://www.pyimagesearch.com/2017/12/11/image-classification-with-keras-and-deep-learning/

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*. https://doi.org/10.1038/323533a0

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel Distributed Processing*. *Processing*.

Russell, S., & Norvig, P. (2010). *Artificial Intelligence A Modern Approach Third Edition*. *Pearson*. https://doi.org/10.1017/S0269888900007724

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*. https://doi.org/10.1016/j.neunet.2014.09.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. https://doi.org/10.1109/ICCV.2017.74

Simeone, O. (2018). A Very Brief Introduction to Machine Learning with Applications to Communication Systems. *IEEE Transactions on Cognitive Communications and Networking*. https://doi.org/10.1109/TCCN.2018.2881442

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.2017*. *UCL,Computer Science Department, Reinforcement Learning Lectures*. https://doi.org/10.1109/TNN.1998.712192

Tanaka, J. W., & Farah, M. J. (1993). Parts and Wholes in Face Recognition. *The Quarterly Journal of Experimental Psychology Section A*. https://doi.org/10.1080/14640749308401045

Tanaka, J. W., & Pierce, L. J. (2009). The neural plasticity of other-race face recognition. *Cognitive, Affective and Behavioral Neuroscience*. https://doi.org/10.3758/CABN.9.1.122

Towler, J., Gosling, A., Duchaine, B., & Eimer, M. (2012). The face-sensitive N170 component in developmental prosopagnosia. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2012.10.017

Turano, M. T., Marzi, T., & Viggiano, M. P. (2016). Individual differences in face processing captured by ERPs. *International Journal of Psychophysiology*. https://doi.org/10.1016/j.ijpsycho.2015.12.009

Turkish Statistical Institute. (2018a). *TurkStat, Population Projections, 2018-2080. Population by age group and sex, 2018, 2023, 2040, 2060, 2080*. Retrieved from http://www.tuik.gov.tr/PreTabloArama.do

Turkish Statistical Institute. (2018b). *TurkStat, Survey on Information and Communication Technology (ICT) Usage Survey in Households and by Individuals, 2004-2018. Individuals using the computer and Internet in the last 3 months by age groups and sex.* Retrieved from http://www.tuik.gov.tr/PreTabloArama.do

Vapnik, V. (1995). The nature of statistical learning theory - excepts. *The Nature of Statistical Learning Theory*. https://doi.org/10.1007/978-1-4757-2440-0

Wiese, H., Komes, J., & Schweinberger, S. R. (2012). Daily-life contact affects the own-age bias and neural correlates of face memory in elderly participants.

Neuropsychologia. https://doi.org/10.1016/j.neuropsychologia.2012.09.022

APPENDICES

APPENDIX A

SOURCE CODE FOR THE COMPUTING BASED EXPERIMENTS

Source code and full image outputs for the computing based experiments, ANN and SVM, are published in https://github.com/caggursoy/Cross-Age-Effect.

APPENDIX B

HUMAN SUBJECTS ETHICS COMMITTEE APPLICATION RESULT

As the thesis involved an experiment on human subjects, the researchers were obliged to obtain a permission from METU Human Subjects Ethics Committee. In order to get a permission from the committee, an application was presented to the committee. The application was approved with the protocol number **2018-FEN-066**, and the approval document can be seen on the next page.

UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
uygam@metu.edu.tr

Sayı: 28620816/ 645

11 ARALIK 2018

Konu: Değerlendirme Sonucu

Gönderen: ODTÜ İnsan Araştırmaları Etik Kurulu (İAEK)

İlgi: İnsan Araştırmaları Etik Kurulu Başvurusu

Sayın Dr.Öğretim Üyesi Murat Perit ÇAKIR

Danışmanlığını yaptığınız Necati Çağatay GÜRSOY'un **"Sinir Ağları'nda Yaşlar-Arası Etki: Sinir Ağları Kullanılarak Yüzsel Yaş Teşhisi'ndeki Eğilimin Ölçülmesi Hakkında Bir Çalışma"** başlıklı araştırması İnsan Araştırmaları Etik Kurulu tarafından uygun görülerek gerekli onay **2018-FEN-066** protokol numarası ile protokol numarası ile araştırma yapması onaylanmıştır.

Saygılarımla bilgilerinize sunarım.

Prof. Dr. Tülin GENÇÖZ
Başkan

Prof. Dr. Ayhan SOL
Üye

Prof. Dr. Ayhan Gürbüz DEMİR
Üye

Prof.Dr. Yaşar KONDAKÇI (4.)
Üye

Doç. Dr. Ali Emre TURGUT
Üye

Doç. Dr. Emre SELÇUK
Üye

Doç.Dr. Üyesi Pınar KAYGAN
Üye

APPENDIX C

EXAMPLE CONSENT FORM

ARAŞTIRMAYA GÖNÜLLÜ KATILIM FORMU

Bu araştırma, ODTÜ Bilişsel Bilimler Bölümü Yüksek Lisans öğrencisi Necati Çağatay Gürsoy tarafından, ODTÜ Bilişsel Bilimler bölümü öğretim elemanlarından Yrd. Doç. Dr. Murat Perit Çakır danışmanlığında yürütülen bir çalışmadır. Bu form sizi araştırma koşulları hakkında bilgilendirmek için hazırlanmıştır.

Çalışmanın Amacı Nedir?

Araştırmanın amacı, kişilerin insan yüzlerinden yaş tahmini yaparken kendi yaşlarının bu tahminlerde ne gibi bir rol oynadığının incelenmesidir.

Bize Nasıl Yardımcı Olmanızı İsteyeceğiz?

Araştırmaya katılmayı kabul ederseniz, sizden beklediğimiz, göreceğiniz insan yüzlerinin yaşlarını tahmin ederek ilgili kutucuğa bu yaş tahminlerini girmenizdir.

Sizden Topladığımız Bilgileri Nasıl Kullanacağız?

Araştırmaya katılımınız tamamen gönüllülük temelinde olmalıdır. Çalışmada sizden yalnızca isim, soyisim, yaş ve cinsiyet bilgileri ile göreceğiniz yüzlere yapacağınız yaş tahminini alacağız. Tahminleriniz ve verdiğiniz bilgiler tamamıyla gizli tutulacak ve sadece araştırmalar tarafından değerlendirilecektir. Siz katılımcılardan alacağımız veriler toplu hâlde değerlendirilecek ve bilimsel yayımlarda kullanılacaktır. Yalnızca her bir katılımcının tahmin verisini bir diğerinden ayırmak için kişisel bilgileriniz kaydedileceğinden, yaş harici diğer bilgileriniz kesinlikle yayımlanmayacaktır.

Katılımınızla ilgili bilmeniz gerekenler:

Deney, kesinlikle kişisel rahatsızlık verecek uygulamalar içermemektedir. Katılımınız sırasında kendinizi herhangi bir sebepten rahatsız hissetmeniz durumunda deneyden çıkmakta serbestsiniz. Bunun için web sayfasından çıkış yapmanız yeterli olacaktır.

Araştırmayla ilgili daha fazla bilgi almak isterseniz:

Deney sonunda, bu çalışmayla ilgili sorularınızın cevaplanması için gerekli linkler size verilecektir. Bu çalışmaya katıldığınız için şimdiden teşekkür ederiz. Araştırma hakkında daha fazla bilgi almak için Bilişsel Bilimler Bölümü öğretim üyelerinden Dr.Öğr.Üyesi Murat Perit Çakır (E-posta: perit@metu.edu.tr) ya da Bilişsel Bilimler Bölümü yüksek lisans öğrencisi Necati Çağatay Gürsoy (E-posta: cagatay.gursoy@metu.edu.tr) ile iletişim kurabilirsiniz.

Yukarıdaki bilgileri okudum ve bu çalışmaya tamamen gönüllü olarak katılıyorum.

İsim Soyad Tarih Onay

----/----/-----

PARTICIPANT CONSENT FORM

This research is conducted by Necati Çağatay Gürsoy, who is a Master's student at METU Cognitive Sciences Department, and Asst. Prof. Murat Perit Çakır, who is a faculty member at METU Cognitive Sciences Department. This form is prepared in order to inform you about the details of the research and the test that you will participate.

Aim of the Research

The aim of the research is to investigate the effect of individuals' own ages on guessing the age from any given human facial image.

How Are You Going to Contribute to the Research?

If you choose the participate, what we expect from you is to guess the ages that you will observe during the run by typing your guesses into the boxes provided.

How the Data Collected Will Be Processed?

Your participation is purely voluntary. During the run; your name, surname, age, gender and your guesses will be collected from you. Your guesses and the information provided will be kept private and only the researchers can have access to any data provided. The collected data will be processed as a whole and it will only be used in scientific publications. Your personal information, apart from your age, will not be published in any case; as it will be collected just to distinguish the guesses of each participant.

What You Need to Know About the Test

The test does not contain any application that would imply any discomfort to participants. In any case of discomfort, feel free to terminate the research. In order to terminate, closing the web page would be enough.

For Further Information About the Research

At the end of the test, links will be provided in order you to ask questions about the research. Thank you for your participation and invaluable effort. For further information, you can contact Asst. Prof. Murat Perit Çakır (perit@metu.edu.tr), who is a faculty member at METU Cognitive Sciences Department and Necati Çağatay Gürsoy (cagatay.gursoy@metu.edu.tr), who is a Master's student at METU Cognitive Sciences Department.

I have read all the information provided above and I'm participating in this research voluntarily.

Name and Surname Date I Approve

----/----/-----

APPENDIX D

PAIRWISE COMPARISON TABLE OF BEHAVIOURAL EXPERIMENT DATA

Pairwise Comparisons

Dependent Variable: Error (Guess - Actual)

(I) Objective Distance (Self Age - Real Age) (Binned)	(J) Objective Distance (Self Age - Real Age) (Binned)	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
< -64	(-63, -36)	-5.956*	1.386	.000	-10.026	-1.886
	(-35, -9)	-9.835*	1.378	.000	-13.881	-5.789
	(-8, 18)	-13.577*	1.376	.000	-17.619	-9.535
	(19, 45)	-13.809*	1.385	.000	-17.878	-9.740
	> 46	-15.486*	2.037	.000	-21.468	-9.505
(-63, -36)	< -64	5.956*	1.386	.000	1.886	10.026
	(-35, -9)	-3.879*	.338	.000	-4.872	-2.887
	(-8, 18)	-7.621*	.333	.000	-8.599	-6.643
	(19, 45)	-7.853*	.369	.000	-8.936	-6.770
	> 46	-9.531*	1.538	.000	-14.047	-5.014
(-35, -9)	< -64	9.835*	1.378	.000	5.789	13.881
	(-63, -36)	3.879*	.338	.000	2.887	4.872
	(-8, 18)	-3.742*	.297	.000	-4.613	-2.870
	(19, 45)	-3.974*	.336	.000	-4.961	-2.986
	> 46	-5.651*	1.530	.003	-10.146	-1.157
(-8, 18)	< -64	13.577*	1.376	.000	9.535	17.619
	(-63, -36)	7.621*	.333	.000	6.643	8.599
	(-35, -9)	3.742*	.297	.000	2.870	4.613
	(19, 45)	-.232	.331	1.000	-1.205	.741
	> 46	-1.910	1.529	1.000	-6.401	2.582
(19, 45)	< -64	13.809*	1.385	.000	9.740	17.878
	(-63, -36)	7.853*	.369	.000	6.770	8.936
	(-35, -9)	3.974*	.336	.000	2.986	4.961
	(-8, 18)	.232	.331	1.000	-.741	1.205
	> 46	-1.678	1.537	1.000	-6.193	2.838
> 46	< -64	15.486*	2.037	.000	9.505	21.468

	(-63, -36)	9.531*	1.538	.000	5.014	14.047
	(-35, -9)	5.651*	1.530	.003	1.157	10.146
	(-8, 18)	1.910	1.529	1.000	-2.582	6.401
	(19, 45)	1.678	1.537	1.000	-2.838	6.193

Based on estimated marginal means

*. The mean difference is significant at the .05 level.

b. Adjustment for multiple comparisons: Bonferroni.