2021

Human Age and Gender Classification using Convolutional Neural Networks

Eamon Kelliher
Technological University Dublin

Recommended Citation

Kelliher, E. (2021). Human Age and Gender Classification using Convolutional Neural Networks. Technological University Dublin. DOI: 10.21427/VBE0-WX80

Human Age and Gender Classification using Convolutional Neural Networks

OLLSCOIL TEICNEOLAÍOCHTA
BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL
UNIVERSITY DUBLIN

Eamon Kelliher

M.Sc. in Computing (Data Science)

2021

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Science), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Eamon Kelliher

Date: 16/06/21

Abstract

In a world relying ever more on human classification, this papers aims to improve on age and gender image classification through the use of Convolutional Neural Networks (CNN). Age and gender classification has become a popular area of study in the past number of years however there are still improvements to be made, particularly in the area of age classification. This research paper aims to test the currently accepted fact that CNN models are the superior model type for image classification by comparing CNN performance against Support Vector Machine performance on the same dataset. Using the Adience image classification dataset, this research also focuses on the implementation of data augmentation techniques, some more novel than others, as a means of improving CNN performance. In terms of standard popular methods of augmentation, image mirroring and image rotation were applied. As well as these, a more novel approach to augmentation was applied to the area of age classification. This technique was completed using Faceapp, an AI image editor in the form of a mobile application. This application allows for the placement of ”filters” on images of human beings in order to alter their appearance. The results of the data augmented models were superior to that of the standard CNN models with gender classification improving by 2.6% while age classification improved by 7.1%. The results of this research establish the potential for further improvements through the inclusion of more augmentation techniques or through the use of more filter types provided in the Faceapp application.

Keywords: Convolutional Neural Network, Support Vector Machine, Data Augmentation, Classification

Acknowledgments

Sincere thanks to my supervisor, Prof. Vincent McGrady. I am truly grateful for your continuous support, guidance and advice throughout this past six months. It was through your mentorship that I was able to push myself outside my comfort zone and complete a thesis in an area of strong interest to me.

I also wish to thank my friends and family, particularly my parents for always supporting and reassuring me that everything would work out. This accomplishment would not have been possible without them.

Finally I would like to thank my employer, Allied Irish Banks, and more directly the Data & Analytics department for providing support throughout my time in Technological University Dublin.

Contents

List of Figures

List of Tables

List of Acronyms

CNN	Convolutional Neural Network
SVM	Support Vector Machine
ELSDSR	English Language Speech Database for Speaker Recognition
DES	Danish Emotional Speech database
PNN	Probabilistic Neural Networks
K-NN	K-Nearest Neighbor
GMM	Gaussian Mixture Model
RBF	Radial Basis Function
RGB	Red, green, blue
EEG	Electroencephalogram
LSTM	Long Short-Term Memory
BLSTM	Bidirectional Long Short-Term Memory
HyperBF	Hyper Basis Function
FERET	Face Recognition Technology
LBP	Local Binary Patterns
DCT	Discrete Cosine Transform
STL	Single Task Learning
MTL	Multi Task Learning
DMTL	Deep Multi Task Learning
IMDb	Internet Movie Database
ReLu	Rectified Linear Unit
RAM	Random Access Memory

Chapter 1

Introduction

1.1 Background

Age and gender classification is a very topical subject area, particularly in this day and age as biometric information is being used for an increasing number of purposes. Such purposes can include medical screening checks, biometric authentication for accessing systems, or for law enforcement purposes. With the ever-increasing use of age and gender classification in our daily lives, there is a resulting reliance on models and applications that are capable of performing this task with more accuracy, precision, and reliability. As humans we have an innate ability to accurately identify another human's age and gender based on our built up knowledge of what constitutes a male versus female and how young versus old people differ in appearance. And while as humans we are capable of this without much effort, it is significantly more difficult for machines, specifically computers, to do this same activity. Machine learning and deep learning techniques have previously been applied to the area of age and gender classification with changes in approach and improvements in results seen over the years. Covolutional Neural Networks (CNNs) and Support Vector Machines (SVMs) have been popular model choices to perform this task, particularly in the case of age and gender image classification. Therefore this research will focus on the use of CNNs and SVMs to complete an age and gender image classification task while implementing new novel techniques to attempt to improve on past research.

The research question is defined as:

”To what extent can convolutional neural networks classify the human characteristics of age and gender through the use of human image profiling?”

1.2 Research Project/problem

Convolutional neural networks are the state of the art method used to classify features from an image. A variety of multi-layer CNNs have been created in previous research to accurately classify the age and gender of subjects from images. Previous research has shown a high degree of success when classifying the gender of the subject in an image. Unfortunately it is not so simple for classifying age. One of the key problems faced by researchers with this topic is the idea of the *one-off* problem. The one-off problem is specific to the classification of age. As previous research has identified that classifying a person’s exact age is extremely difficult and often leads to undesirable results, many researchers have resorted to binning ages in groups in order to achieve a more accurate level of classification. It is with these age groupings that the one-off problem occurs. The one-off problem occurs when the CNN, or any model, classifies a subject to be in the age group exactly one above or one below the actual age group the subject is a member of. And prior research has shown that the accuracy of classifying a subject to their actual age group versus one-off varies significantly (Levi & Hassncer, 2015). An aim of this research will be to not only improve the standard age classification accuracy but also to improve the one-off accuracy of that achieved by Levi & Hassncer (2015). An improvement in both metrics will result in a more accurate and higher performing CNN model than that of Levi & Hassncer (2015). Previous research has found that the leading cause for this one-off problem is the limited amount of data acquired and used to train and test the CCN with. Because of this, data augmentation techniques will be implemented within this research to boost the number of samples in the dataset. In particular, image mirroring, image rotation, and the use of the Faceapp mobile application will be used to augment the dataset.

As well as the one-off problem associated with age classification, small strides are also still possible for improving gender classification. The once considered state-of-the-art research for performing gender classification was created by Levi & Hassncer (2015) and managed to achieve gender accuracy of 86.8%. And while the main focus of this research is to improve on the age accuracy result achieved by Levi & Hassncer (2015) which was 50.7%, it is believed that the gender classification result can also be improved. This will be attempted by providing more data to the CNN during the training process using the image mirroring and image rotation mentioned previously. In terms of the datasets available for use for age and gender classification, the issue of dataset bias is a common one in this area with many datasets having significantly more subjects in certain age groups that others (Levi & Hassncer, 2015). In particular with the Adience dataset, subjects between the ages of 25-32 are more common than other age groups leading to bias within CNNs and a tendency for models to incorrectly classify a subject in this age bracket. The hope is that through the use of data augmentation techniques to provide the CNN with more a more balanced and larger dataset to learn from, that this one-off problem will reduce significantly and that the model will be able to perform age and gender classification at a far greater level in terms of model accuracy.

1.3 Overview of methods to overcome research problem

This one-off problem has been seen in a significant number of research papers that aim to tackle the issue of age classification of humans. CNNs typically require a substantial amount of data in order to become successful models [1]. Unfortunately it is often computationally or monetarily expensive to acquire the large amounts of data required to create successful CNN models. Given the nature of this issue, the one-off problem has been a persistent one over the years. As a result, the novel approach that will be used in this research paper in an attempt to overcome this revolves around

[1] 4 Reasons Why Deep Learning and Neural Networks Aren't Always the Right Choice, 2021

the use of a number of data augmentation techniques. One in particular is only a recent possibility due to the creation of a mobile application which will be discussed further below. The initial data augmentation techniques that will be used are not exactly novel however are seldom seen in research papers aiming to classify the age and gender of humans.

This first augmentation technique is image mirroring which in essence involves flipping the images and saving these as copies of the original. This is a simple yet effective means of increasing the number of images in the dataset and does not require a significant workload. Images can be mirrored vertically, horizontally, and mirrored in both axes resulting in what was originally one image now becoming four. For this research specifically however, only horizontal image flipping will be applied. The reason for this is that in the context of human image classification, it does not make sense to use upside-down images of humans for classification purposes. In other image classification challenges, such as those that use aerial images of land or the sea for examples, vertical image flipping would make more sense as a data augmentation technique. If vertical flipping was applied for this human classification task, the model would require even more images than before to be able distinguish between a male or female or between people in different age brackets in an upside-down setting. The main reason for this being that a picture of a person upside down would look nothing like the same picture the right way up, particularly to a CNN model.

Image rotation is another useful method of data augmentation that will be applied in this research. As will be discussed in further detail in the Literature Review section, image rotation has seen plenty of success in helping to increase the accuracy results of CNNs. Image rotation differs slightly to image mirroring in that instead of creating a new image that is the opposite way around, image rotations turns the image a specified number of degrees. Similarly to the reason why vertical image mirroring should not be used for human image classification, it is important not to rotate the image to the point that the subject of the image is no longer somewhat upright. Small rotation values between 5 and 10 degrees are appropriate for this classification task. The idea behind rotating the image is that by moving the pixels around, the image looks like a brand new one to a CNN, although it may look very similar to the human eye. Again, this is a simple augmentation technique that has shown promise in previous research and appeared to be a great choice for this research.

The next and more novel data augmentation technique that will be applied to the images is going to be through the use of the FaceApp mobile application. This application is an AI photo and video editing service that allows users to upload images of humans and then apply a number of different ”filters” to the image to alter the appearance of the person in the image. Of all the applications capable of altering images of humans, Faceapp was chosen due to the fact that it was a free service to use, while allowing for the easy exporting of the altered images. Other applications that have the ability to alter images in a similar manner tend to either charge a premium fee for their service or else do not provide the ease of image exporting that Faceapp does. Faceapp also has an excellent reputation in the image AI editing field with excellent reviews on both the Apple App store and Google Play store. Finally, having tried and tested a number of applications for this process, the filters applied by Faceapp tended to look more natural and authentic than the filters applied by other applications.

The purpose of this application is to both augment the dataset while also making the dataset more balanced overall. As mentioned previously, the dataset that is being used in this research has an imbalance towards subjects between the ages of 25-32 with a larger number of subjects falling into this age bracket than any other. The idea is to take the images of subjects in this age bracket and to apply an ageing filter to them to make the subjects of the images appear significantly older. The age increase caused by the filter then gives a new image of a person who would fall into the age bracket of 60-100. The methods employed by Faceapp to make a person appear older is mainly through the introduction of wrinkles, through the greying and slight receding of hair, and through making the person’s skin appear more elastic or ”saggy”. All in all, Faceapp does a convincing job of making a person appear to be significantly older than they are. By adding this ageing filter to approximately 500 images of people between the age of 25-32, the dataset will still contain the original image but will now also have a second image of the same person but looking significantly older. Theoretically this will result in a more balanced dataset as the number of elderly participants will increase potentially resulting in an overall improvement in model accuracy and loss and a reduction of the one-off problem. Unfortunately, the process of applying the ageing filter to images is a manual one and therefore the filter was applied one by one to each of the chosen images before being exported one at a time to include them in the dataset.

1.4 Research Objectives

The purpose of this research is to create CNNs with the ability to accurately classify a human's age and gender based on a facial image of that person. The desired outcome is models that would be capable of taking any particular image of a humans face and classifying the age and gender of this human with a high degree of confidence and accuracy. With this in mind, to further support this research, separate SVM models will also be created as a means of comparison to determine the optimal model solution to complete the task of image classification. The SVMs will be used to classify the age and gender of humans using the same image dataset resulting in directly comparable results and a deeper insight into the superior model type for completing such a task.

Further to this, a secondary objective will be gain further understanding in the area of data augmentation and the impact it can have in a human age and gender image classification task. As will be discussed in the Literature Review section, data augmentation can have both positive and negative impacts on model performance. And while in the vast majority of cases data augmentation leads to more successful models, this research will focus on understanding the impact this technique has on age and gender classification using the Adience dataset.

1.5 Research Methodologies

The research methodologies employed in this research will be discussed under the following 4 headings - by type, by objective, by form, and by reasoning.

1.5.1 By Type - Primary vs Secondary Research

Primary research involves the direct collection of data by researchers. In this scenario, the researchers are creating the data through their means of collection and it is this data that is used to carry out the research. Secondary research on the other hand involves relying of pre-collected data or data that already exists. The researcher does not need to do any data collection themselves in this scenario and can rely solely on data that has been collected by somebody else.

This research employs both a primary and secondary research methodology approach. This paper is mainly secondary research focused as the majority of the data used was previously gathered and labelled by researchers from the Open University of Isreal. However a small element of primary research was undertaken through the use of the Faceapp augmentation technique. By using Faceapp, new image data was generated through the use of images from the pre-existing Adience dataset.

1.5.2 By Objective - Quantitative Research

Quantitative research is typically expressed in numbers and graphs and is a research methodology used to confirm or test theories and assumptions.

With this is mind, this research paper is considered quantitative research. The purpose of this research is to test different model types against one another to confirm the superior model type for an image classification task. As well as that, this research is also looking to test the impact data augmentation techniques have on the results of model performance.

1.5.3 By Form - Exploratory vs Constructive vs Empirical

Research is considered exploratory research when the problem the research is looking to solve has not been clearly defined. Exploratory research aids in determining the best research design and data collection methods. Constructive research on the other hand is often applied in the area of computer science. This type of approach demands a form of validation that doesn't need to be quite as empirically based as in other types of research like exploratory research.[2]. Finally, empirical research typically involves the creation of a hypothesis which will then be tested using the required experiment.

Given the descriptions above, this research paper is an empirical form of research due to the fact that multiple hypotheses have been stated with each of them being tested by the outputs of the CNN and SVM models created as part of this research.

1.5.4 By Reasoning - Deductive Research

The key difference between inductive and deductive research is that inductive research has the goal of developing a theory whereas deductive research is used when testing an existing theory. Deductive research begins looking at a general rule and by the end has established a guaranteed specific conclusion.

Given the differences in the two, this research falls under deductive research. This is because the research started with a general theory about whether a human's age and gender could be classified from an image. From there a number of hypotheses were developed before finally data was gathered and used to create models to determine whether the hypotheses could be accepted or rejected.

1.6 Scope and Limitations

The scope of this research is to develop both a CNN and SVM using the Adience dataset to classify the age and gender of subjects in images.

[2] Constructive Research — Psychology Wiki — Fandom, n.d.

One of the main limitations with this research involves the filters that the Faceapp application offers. The filters do not allow for the ageing on humans by a particular number of years. Instead the filters that exist within the application are generic such 'Child', 'Teen', or 'Old', making the subject of the image look how the name of the filter describes. Because of the nature of the age label groupings within the Adience dataset (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-100), it was only possible to use the 'Old' Faceapp filter on the images. The reason for this being that the 'Old' filter was the only filter where one could be confident of the age grouping the subject of the image now belonged to (60-100). For example if one were to use the 'Child' filter on an image it would result in subjective decisions being made as to whether the subject in the image now looked between the ages of 4-6 or 8-13. This subjectivity meant that only the 'Old' filter was used in this research.

Another limitation revolves around the manual nature of applying the Faceapp filters to images. Unfortunately there is no bulk upload service provided by Faceapp meaning that the 'Old' filter was applied to images one by one. Due to this time pressure associated with this research project, this manual nature of Faceapp meant that approximately 500 images had the 'Old' Faceapp filter applied. Had more time been available for this research, the filter would have been applied to more images.

1.7 Document Outline

In this paper, an attempt is made to further improve on previous research in age and gender classification. Using the Adience dataset, a very popular dataset designed specifically with age and gender classification in mind, a CNN was developed to classify the age and gender of humans from images. This paper contains a Literature Review to discuss previous work in this area, along with similar research that uses slightly different approaches to solve the same problem. Gaps in previous research have been identified and will be discussed further throughout this paper with the core gaps surrounding the issues of the one-off problem and unbalanced datasets. Following the discussion of previous research, the next section focuses on the methodology used to complete the task. Here the design of the CNN and SVM architectures will be discussed along with more detail on the data augmentation techniques and the dataset used to complete the task. Finally, the results will be analysed and a decision will be made on whether the data augmentation techniques resulted in an increase in model performance over a model that was trained without any data augmentation techniques. The results will also be compared against those achieved by Levi & Hassncer (2015), a paper that was once quoted as being the state-of-the-art in this area. In this section, comparisons will also be made against the performance of an SVM for the same classification task to help determine the best model type for age and gender classification. The paper will then be rounded off with the conclusion along with brief details on the potential of future work.

Chapter 2

Review of existing literature

2.1 Literature Review

The gold standard state-of-the-art approach for image processing and classifying objects from images is the use of convolutional neural networks. This is evident from the previous research completed in the area of classifying the age and gender of humans from images. Various CNN architectures have been used in the past with different levels of results achieved however it is still the dominant model to use to carry out such a classification task. The different approaches used by other researchers to complete this image classification problem will be discussed below. As well that similar CNN approaches used for solving different problems will also be covered. The purpose of this section is to show the popularity of CNNs for image classification while also showing how the gaps this research aims to fill were discovered.

2.1.1 Age and Gender classification using alternative approaches

Convolutional neural networks have overtime become the state-of-the-art means of performing image classification. Whether it be binary or multi-class image classification, CNNs have become the model of choice amongst researchers who work in this field. However this does not mean that other approaches are still not being considered and used today. An example of an alternative approach can be seen taken in the research completed by Qawaqneh et al. (2017). Qawaqneh et al. (2017) took the approach of using 2 deep neural networks to classify age and gender. This research expanded on just the use of facial images and also employed speech audio to complete the classification process. The researchers created a new cost function to fine-tune the deep neural networks jointly with the cost function being evaluated using speech utterances and unconstrained face images for an age and gender classification task. The freely available databases, the Age-Annotated Database of German Telephone Speech database and the Adience database were used to measure the performance of the system. By training the deep neural networks using this new cost function, it resulted in improved classification accuracy and reduced the issue of overfitting for existing speech-based and image-based systems. Overall this research resulted in an accuracy of 56.06% for seven speaker classes and 63.78% from the Adience database.

As seen with the research completed by Qawaqneh et al. (2017), images are not always the only choice of data used when researchers are looking to complete age or gender classification. This fact can also be applied to the research completed by Sedaaghi, M. H. (2009) which did not implement the use of images at all. Rather than focusing on images to classify age and gender of humans, research conducted by Sedaaghi, M. H. (2009) used speech signals to perform this task. Using the Danish Emotional Speech database (DES) and English Language Speech Database for Speaker Recognition (ELSDSR) database, a number of different algorithms were applied to help determine the best classifier when it comes to such a task. In particular probabilistic neural networks (PNNs), support vector machines (SVMs), K nearest neighbor (K-NN) and Gaussian mixture model (GMM), as different classifiers, were implemented and empirically compared to determine the best classifier for gender and age classification when speech signal is processed. The SVM achieved the best results for gender classification with an impressive 94.83% while an accuracy of 88.38% was achieved by the PNN for age classification.

Support Vector Machines have shown great promise in the area of classification with an example discussed above where the best model created by Sedaaghi, M. H. (2009) for their classification task was an SVM. And while it was speech classification performed by Sedaaghi, M. H. (2009) that achieved their excellent results, SVMs have also proved to be excellent model choices in the image classification space too. An example of this can be seen with research completed by Fazl-Ersi et al. (2014). Fazl-Ersi et al. (2014) aimed to classify age and gender using SVMs with RBF kernels. This study created a model built primarily on the Gallagher dataset, a dataset made up of mostly real-world images making the classification process that much more difficult. Unlike the majority of other solutions to this problem that tend to focus on a single visual descriptor which encodes only a particular characteristic of the image regions, this research incorporates multiple feature types therefore taking advantage of various sources of information. To improve further on this, only the regions that best separated the facial images into different demographic classes (with respect to age and gender) are used to create the facial representations. Due to the use of this technique, the output of the model with respect to classification and recognition accuracy improved greatly. The final result of this model achieved a peak accuracy of 91.59% for gender classification and 63.01% for age classification.

Often times image classification models can require the use of the entire image and the content of the entire image, taking in multiple feature sources to perform the classification task to a high level of success. However other examples of research have shown that very specific and focused images can also yield fantastic results for image classification tasks. For example, Rattani et al. (2018) proposed a solution for gender classification using ocular images taken on a smart phone. The use of ocular images in this case is very specific and means that the learning model can only use very limited image data information to complete the required classification task. In the context of smartphone devices, gender information has been used to enhance the accuracy of the integrated biometric authentication and mobile healthcare system. The novelty of the approach used in this research was the use of RGB ocular images to classify human gender which had not been used in prior studies. Specifically, pre-trained CNN architectures were used in order to classify gender. To further improve the results of this research, multi-classifier fusion was used. The results of this research were compared with humans ability to classify gender to establish how effective the results were. The results of the model were also compared against off-the-shelf-texture descriptors as a means of reference to how successful the CNN model was. The results varied depending on the mobile device the ocular image was photographed on with accuracies ranging from 83.3% to 87.6%.

As mentioned previously, researchers have used data such as speech signals and speech audio to complete their respective classification tasks. Qawaqneh et al. (2017) and Sedaaghi, M. H. (2009) are examples of research in which speech was used to a very effective level for the completion of age and gender classification. And while CNNs are the most popular method for classifying age and gender of humans, the use of other non-deep learning methods are also on the rise. Kaushik et al. (2019) took a different and far more advanced approach again to completing their age and gender classification. Research completed by Kaushik et al. (2019) focused on the use of brain-computer interfaces to perform such a task. Through the use of an EEG recording device, this research recorded the cerebral activities of 60 subjects in a relaxed position with closed eyes. A deep BLSTM-LSTM network was used to construct a hybrid learning framework for the analysis. And while this type of research is not specifically machine learning focused, accuracies of 93.7% and 97.5% for age and gender classification respectively show that CNNs are not the only means to perform such a task. However this research is far more limited in terms of the number of subjects that can be analysed as actual brain signals of the subjects need to be monitored rather than just the use of an image of the subject. This therefore makes this research less applicable in an everyday situation.

Typically in the world of CNNs and image classification, a substantial amount of data is required to create and train a model that is capable of taking new and previously unseen data and classifying this data with a high degree of accuracy. This statement is not always true for different model types and techniques that have been applied to the area of age and gender classification. As seen above in the research completed by Kaushik et al. (2019) in which results of over 90% were achieved for age and gender classification with only the use of 60 patients data required, a similar scenario and result can be found in work completed by Kwon & da Vitoria Lobo (1999). This work was one the earlier studies in this area and focused specifically on age classification from facial images. At the time this research was completed, the previous works in this area mainly focused on only 3 age categories from which to classify a person into - babies, young adults, and senior adults. These same categories were used in this analysis. This paper was one of the first of it's kind to successfully extract and use natural wrinkles to classify a humans age. Using computations based on cranio-facial development theory and wrinkle analysis, the algorithm achieved an accuracy of 100% for classifying humans within one of the 3 categories. However a major drawback of this research is that only 15 images were used to test the model as that was the maximum number of images available in which wrinkles could be analysed.

The examples above are not the only examples of research discovered during the process of this Literature Review that used a limited data size when developing a classification model. And while the examples discussed previously still resulted in models that had very high levels of test accuracy, there are examples of research where the lack of data has resulted in models with below average performance compared to other research models also performing age and gender classification. One such example is a paper completed by Poggio et al. (1995). Research completed by Poggio et al. (1995) focused on gender classification through the use of HyperBF networks which were created from a set of of geometric features that were extracted from digitalised photos of humans in a frontal pose. This resulted in a database of 20 males and 20 females which was used to train the HyperBF network. After applying this database in the training process, the resulting model was able to perform gender classification with an accuracy of 79% on any new images passed to the network. These results were compared to human performance on the same set of grey-scale images which achieved an average result of 90%. And while the results of the network could not match that of classifications performed by humans, an interesting finding discovered was that the HyperBF technique finds the relative weights of the different features and converges to prototypes of the male and female face that seem to exaggerate their difference, somewhat like caricatures do.

As mentioned prior in this related work section, SVMs are also another popular method use for image classification and research completed by Moghaddam, B., & Yang, M. H. (2000) also focused on gender classification using SVMs. SVMs were a very popular algorithm to perform such a task in the year 2000 and, in this research, SVMs outperformed other traditional classifiers such as linear, quadratic, Fisher linear discriminant, nearest neighbour as well as other more modern techniques of the time such as radial basis function at classifying gender from images. Using the FERET face database that consists of 1,755 low resolution "thumbnail" facial images, the SVM achieved the lowest error rate of 3.4%. The SVM was also tested on higher resolution images where it also outperformed the other comparable models.

A point of note that has cropped up time and time again with related work in the age and gender classification space is that age classification tends to be the more difficult of the two tasks to create an effective model for. This is to be expected as gender is a binary problem and age classification is a multi-class problem. This has resulted in researchers having to take different approaches to solve the issue. The most common approach is to bin the ages into groups of approximately 5 years. This allows for a greater scope from which to classify an age to a human. A slightly different approach to this age binning was taken in the research completed by Günay, A., & NabIyev, V. v. (2008). This research focused on automatic age classification through the use of local binary patterns (LBP). LBPs are fundamental properties of local image texture and the occurrence histogram of these patterns is an effective texture feature for face description. The results of this research was a model that could classify age with an accuracy of 80%. However this result was not particularly fantastic when compared to other age classification models. The main reason for this being that this research classified ages into 10 year groupings which is a rather large interval when compared with other research completed in this area.

2.1.2 Convolutional Neural Networks for more than just Age and Gender classification

Age and gender image classification are very popular forms of research when it comes to classifying anything human related. However that is not to say that age and gender are the only areas of interest when it comes to classifying characteristics or demographics of humans. This section will focus on research that went further than just age and gender classification and research focused on different topics of classification altogether. Further to this, this section will be focused solely on research that implemented the use of CNNs in order to perform this task. This will give an idea of the further uses of CNNs and the success they can have in other areas of classification. The first research paper covered in this section takes the idea of age and gender classification to another level with a far more specific category of classification included in the research. Srinivas et al. (2017) created a model with the goal of classifying the age, gender, and fine-grained ethnicity of images of people from East Asia. The dataset used in this research was the Wild East Asian Face Dataset which was considered a new and unique dataset at the time and consisted of labelled facial images of humans from East Asian countries such as Vietnam, Thailand, China, Korea, and Japan. East Asian Amazon Mechanical Turk annotators were used to label the age, gender and fine grain ethnicity attributes to reduce the impact of the "other-race effect" and improve quality of annotations. The specific model used to perform this classification task was a CNN, and two separate architectures were created in the research process. The results of this research were mixed with only gender classification achieving a solid accuracy. Age classification reached an accuracy of 38.04%, gender classification reached an accuracy of 88.02%, and fine-grained ethnicity reached an accuracy of 33.33%. A lack of data augmentation techniques was seen as a key reason for the poor age and fine-grained ethnicity classification scores.

As seen in the research completed by Srinivas et al. (2017), age and gender are not the only areas of interest for researchers working in the area of human image classification. And while the research completed by Srinivas et al. (2017) lacked the dataset size to create a model that could classify fine-grained ethnicity of the Asian population with any great degree of success, other research that went beyond just age and gender classification has seen excellent final results. An example of this can be seen in a paper completed by Dehghan et al. (2017) in which they explored more than just age and gender classification when completing their research. An additional classification task of classifying the emotion a human was feeling in an image was also included. The addition of the emotion classification category is quite unique in this space and adds an extra dimension to the overall study. The researchers created a system known as Sighthound which consisted of several deep CNNs that provide state-of-the-art results on several competitive benchmarks. Sighthound achieved an accuracy of 70.5% for age estimation, 76.1% for emotion classifcation, and 91% for gender classifcation.

The inclusion of an extra category on top of age and gender has shown mixed results thus far with Dehghan et al., 2017 achieving strong results with their emotion classification while Srinivas et al. (2017) created a model that struggled to classify fine-grained ethnicity with any real confidence. However despite the mixed results that can occur when adding in more and more categories to classify, there has been research studies completed that have included significantly more categories for which to classify. Work completed by Ranjan et al (2017) took the idea of facial image analysis and used it to perform numerous activities including face detection, face alignment, pose estimation, gender recognition, smile detection, age estimation and face recognition. All of the aforementioned tasks were completed using an 'all-in-one' convolutional neural network. The method used employed a mulit-task framework that regularizes the shared parameters of CNNs and builds a synergy among different domains and tasks. To take into account all the tasks that were being covered in this research, a number of different datasets were required in order to have all the required labelled information for training the CNN. The combined number of images from each of the datasets was just shy of one million and resulted in a CNN that was proficient at performing each of the tasks set out to do. Compared to previous research called HyperFace, the research performed by Ranjan et al (2017) performed significantly better despite the use of multi-task learning being used in both instances.

The uses of human image classification are many, specifically the uses of facial recognition. In particular, facial recognition is used in our every day lives for activities such as unlocking personal devices like mobile phones or laptops and even used in areas of travel in automated border patrol settings in airports. This area has not been touched thus far in the related work section however it does fall under the area of human image classification. With that in mind an example of facial recognition research was completed by Nagi et al. (2008). It did not focus specifically on age or gender classification in any form but rather more on general facial recognition research. This research used an image-based approach towards artificial intelligence by removing redundant data from face images through image compression using the two-dimensional discrete cosine transform (2D-DCT). Features were extracted from the faces based on skin colour using the DCT. Feature vectors were then created by computing DCT coefficients. Then in order to determine whether a particular individual is present or not present in the image database, a self organising map using an unsupervised learning technique was used to classify DCT-based feature vectors into groups. The database used for this research consisted of 25 images, made up of 5 images of 5 different subjects, each with a different facial expression. The output of this research was a model with a recognition rate of 81.36% which had high-speed processing capabilities and low computational requirements.

Throughout the course of analysing papers focused on the area of age and gender classification, papers were discovered that made use of common datasets typically used for general image classification. A very basic approach to image classification was completed by Rizvi, M. S. Z. (2020). This work was not specific to age or gender classification, rather just standard image classification using the well known datasets MNIST, CIFAR-10, and ImageNet. The MNIST dataset was used in the creation of a CNN that could identify individual hand-written digits with the final output achieving an accuracy of 99% after 10 epochs were used. The CIFAR-10 dataset was used to identify the subjects of image, whether it is a aeroplane, auto mobile, bird, etc. Using a CNN to perform this task, an accuracy of 83% was achieved after 10 epochs. And finally, the ImageNet dataset was used for categorising images with the subjects of the image being random objects with anything from a golf ball to a church included in the dataset. Again using a CNN with 10 epochs, an accuracy of 91% was achieved in this study.

Similar to the work conducted by Rizvi, M. S. Z. (2020), the research performed by 'Uyun, S., & Efendi, T. (2019) took a similar basic approach to image classification. A CNN was created that implemented back-propagation which was then applied to the CIFAR-10 database in an attempt to correctly identify the objects of an image. The resulting model did not perform very well with a maximum validation accuracy of just over 50%.

'Uyun & Efendi (2019) focused on a rather different task through the idea of image classification. Rather than attempting to classify age and gender of humans from images, the researchers instead created an algorithm that could classify the weight and height of a human based on an image. As part of this research a number of algorithms were created, each with a variation in how the weight and height of a human would be calculated. Using an Android smartphone to capture the image of a human, the image was then passed to each of the created algorithms to determine the weight and height of the person in the image. Algorithm 'C' achieved the best results in this research. The approach of this algorithm is to measure the width of the object starting with the height of the image adjusting half of the height of the object in the image. This algorithm achieved results of a deviation value of 1.85% for height and 8.87% for weight. The accuracy rate in determining the ideal body weight has reached 78.7%.

2.1.3 Convolutional Neural Networks for Age/Gender classification

The area of age and gender classification using convolutional neural networks is an ever growing area of popularity which can be seen by the research papers that will be discussed in this section. The growth in popularity can be seen as each of the papers that will be discussed have been completed in the last number of years. Research completed by Levi & Hassncer (2015) was seen as the state-of-the-art research for a long time and is a study often referenced and compared to by research that has been completed since. Levi & Hassncer (2015) focused on the classification of age and gender of humans from facial images using CNNs. Specifically, the researchers created a CNN with 3 convolutonal layers and 2 fully connected layers to classify human age and gender from images. The novel factor of this work was that previous analysis in this area often did not focus on unconstrained imaging conditions. This paper used the Adience dataset which consists of real world images which resulted in a model able to handle the everyday photo while previous work in this area often relied on posed images to create their model. The use of real world images resulted in a far more useful model, capable of classifying age and gender far more effectively than most research seen prior to it. Even with the simple architecture employed, the output of this research was a model that achieved an accuracy of 86.8% for gender classification and 50.7% for classifying age. And while the model performed excellently, the simple nature of the architecture allowed room for more complicated, complete architectures to be created on top of this research, providing the opportunity for further improvements to be made.

The research completed by Levi & Hassncer (2015) discussed above formed the basis for a number of research papers in the area of age and gender classification. This next paper completed by Al-Azzawi, D. S. (2019) was one of those and the researchers took the research completed by Levi & Hassncer (2015) and expanded on it through the introduction of new machine learning methods. Al-Azzawi, D. S. (2019), similarly focused on age and gender classification using CNNs. In this research a new idea was presented by modifying the deep network structure and architecture of previous work they were basing their research off. Further to this, more learning methods were introduced to the research with an aim to boost the final performance of the CNN. These two methods were Single-Task-Learning (STL), which focuses on creating a model to perform a certain task using information directly related to the task at hand, and Deep-Multi-Task-Learning (DMTL), which focuses on using information from the training signals of related tasks. The idea is that by using these training signals and sharing representations between tasks, it can enable a model to better generalise than a model just built using an STL approach. This research was based on the Adience dataset and the introduction of both STL and DMTL showed improvements in performance on previous research that was being used as a comparison. The final output of this research was a CNN capable of classifying gender with an accuracy of 91.34% and predicting age with a Mean Absolute error of 4.00.

Again this next paper took inspiration from the work completed by Levi & Hassncer (2015). However rather than use the Adience dataset, the researchers attempted to complete their task using the IMDb dataset. Agrawal, B., & Dixit, M. (2020) similarly created a convolutional neural network to predict the age and gender of humans from facial images. PCA was applied in this research to reduce the extracted feature dimensions. Unlike Levi & Hassncer (2015) the IMDB-WIKI dataset was used rather than the Adience dataset. The MATLAB platform was used to implement this research.

Varying datasets have been used in the area of age and gender classification. This research paper focuses on the use of the Adience dataset and it is a dataset designed specifically for the task at hand. However another very popular and much larger dataset, the IMDb dataset, is also a very popular choice for such a task. This dataset was the choice of Safak, E., & Bariicc, N. (2018). This research focused on the use of CNNs and their ability to classify the age and gender of humans using facial images. The researchers implemented the Inception V1 convolutional neural network that has been developed by Google for training and has already been trained on the VGGFace2 dataset. This dataset contains 3.31 million images of 9,131 subjects resulting in an average of 362.6 images per subject. Using transfer learning, the CNN was trained using this dataset. The researchers then used this CNN which had employed transfer learning to try and predict the age and gender of humans from the IMDb dataset which consists of 460,723 images. The final results of this model were very impressive with age classification reaching 70.3% accuracy and gender classification reaching 97% accuracy.

While the papers discussed previously in this section focused on both age and gender classification using a variety of different CNN structures and datasets, there are research papers completed that focus on just one single category to classify. This very narrow area of focus has allowed many researchers to create models and achieve results that are extremely accurate and ground-breaking in their field of study. One such example of this is research completed by Antipov et al. (2016) which focused on gender classification using a CNN. These researchers have written a number of papers on this research topic. Using the Labelled Faces in the Wild dataset, a dataset that is considered one of the most challenging set of images for completing this type of task, the final model achieved a ground-breaking result of 97.31%. The dataset used 10 times less images than the previous state-of-the-art model making the process a far more efficient one than this previous state-of-the-art. Further to this, the ensemble model created in this research was designed in a way that both memory requirements and running time are minimised. The impact of this is that the resulting model has the potential to be embedded in mobile devices or in a cloud service for intensive usage on massive image databases.

Further research completed by Antipov et al. (2017) acknowledged that previous research had shown that CNNs had been very effective for human demographic estimation. However it was noted that the approaches used in previous research varied so significantly that there was no fundamental method for ensuring success when undertaking such a challenge as human image classification. In particular this paper focuses on how to choose the optimal CNN architecture and which training-strategy to use. Four key factors were analysed - the target age encoding and loss function, the CNN depth, the need for pre-training, and whether to use mono-task or multi-task for the training strategy. The outcome of this analysis resulted in the creation of three models with the most successful model winning the ChaLearn Apparent Age Estimation Challenge 2016, significantly outperforming the competition.

As previously mentioned in this document, age classification has proven to be a more difficult challenge to complete for researchers than gender classification. Because of this, researchers have had to use techniques such as age binning as a means of classifying humans into a particular age bracket rather that classifying an exact age. In examples where this has occurred, it can be difficult to decide on the number of age brackets to create and how wide an age range should be implemented. This can be seen in research that focused specifically on age classification of humans from grey scale images that was completed by Horng et al. (2001). The grey scale limitation of this research significantly reduces the potential use of this model on typical real world images that would be taken in colour. The age brackets in this research were also fairly limited with only four age groups implemented - baby, young adult, middle-aged adult, and old adult. The limitation to these 4 age groups results in a model that does not sufficiently classify the human in the image with much purpose. The final model achieved an accuracy of 81.58% for classifying a human into one of the 4 specified groups.

The paper completed by Horng et al. (2001) mentioned above was the paper that had the fewest number of age brackets when it comes to papers that focused specifically on age classification. A paper that took this a step further by introducing an additional fifth age group was research completed by Kumar (2020). This paper took a standard approach to classifying the ages of humans from images using CNNs. Taking a labelled dataset with 5 age classes (1-14, 14-25, 25-40, 40-60, and 60+) the final model was able to classify a human into the correct age bracket with an accuracy of 62%. Further improvements that could have been made to this research would have been an increase in dataset size from 23,000 images through the use of data augmentation techniques. It was also noted that a better network architecture could have further improved the results. Analysis of the results showed that the model struggled to accurately classify the ages of subjects that were not facing the camera directly or in cases where images were of poor quality.

Often times it can happen that new and novel approaches to solving deep learning problems can result in worse performance than more traditional approaches to solve the problem. A prime example of this can be seen in this next research paper in which gender classification disimproved after a a newer approach was applied to the data. Deep Multi-Task Learning (DMTL) was used alongside a CNN in an approach used by Ito et al. (2019) in order to predict age and gender of humans from images. The idea behind using DMTL is that classification accuracy should improve by training the network while sharing a part of the network for each task and the loss function. When compared with Single Task Learning (STL), DMTL made it possible to significantly reduce computational time of the prediction. It was found that DMTL improved the accuracy of age estimation while the accuracy of gender estimation was reduced. With DMTL, the CNN classified gender with an accuracy of 93.54% and age estimation with a mean average error of 7.217. While with STL, the CNN achieved an accuracy of 93.86% for gender classification and had a mean average error of 7.327 for age classification.

2.1.4 Data Augmentation techniques to Improve Convolutional Neural Networks Performance

One of the key aspects of this research study is to find a new approach to implementing data augmentation that has not been completed before. This section will be dedicated to analysing novel approaches used in other research areas and the benefits that augmenting a dataset can have on model performance.

Convolutional neural networks have been the model of choice for image classification for the past number of years. However one issue that often crops up when creating a CNN is the issue of data size. Typically, in order to create an effective CNN model, there is a requirement for a large amount of data. This was an issue researched by Han et al. (2018) in which they attempted to create a new image classification method using CNN transfer learning and web data augmentation. The idea behind this research was to find a means of applying superior deep CNNs such as AlexNet, VGG, and ResNet for problems where there were limited data sizes. To overcome the issue of limited data sizes, Han et al. (2018) created a novel two-phase approach involving the use of CNN transfer learning and web data augmentation. Their method allowed for the feature presentation of a pre-trained model such as AlexNet to be transferred to a new task while at the same time the method scraped additional valuable images from the internet that could be included to further bolster the original dataset. This research has created a solution of effectively expanding dataset sizes while also reducing the issue of overfitting often seen with CNNs. By applying their transfer learning and web augmentation methods to previously completed research, Han et al. (2018) saw an improvement in results with the ResNet model outperforming a number of state-of-the-art models when applied to small datasets.

While Han et al. (2018) focused on the use of web scraping images to further augment their datasets, a very different approach was undertaken by Li et al. (2017). These researchers noted that CNNs can provide excellent performance in the area of hyperspectral image classification once there is a significant number of data samples to train the CNN from. Therefore in order to overcome the issue of small datasets, a novel pixel-pair method was proposed to increase the number of training samples. For a testing pixel, pixel-pairs, constructed by combining the center pixel and each of the surrounding pixels, are classified by the trained CNN, and the final label is then determined by a voting strategy. CNNs developed with this pixel pair augmentation strategy using hyperspectral image datasets have shown that the method is capable of outperforming conventional deep learning based methods.

Last year, research completed by Zhong et al. (2020) used a relatively simple but also very effective means of data augmentation when training a CNN in comparison to the work completed by both Han et al. (2018) and Li et al. (2017). The techniques used in this research is called Random Erasing and in essence is used to remove a section of an image and replace it with random pixel values. By applying Random Erasing, images with varying ranges of occlusion are created. By training a CNN with incomplete images it helps to reduce the issue of overfitting often seen in CNN models. Random Erasing can often be used coinciding with image mirroring and and random image cropping techniques that also exist. In this research, the use of Random Erasing was used on very well known datasets such as CIFAR-10, CIFAR-100, and Fashion-MNIST as a mean of understanding if it causes any improvement in image classification. Results from the research indicate that models trained with Random Erasing in place improve significantly compared to those that don't. For

CIFAR-10, Random Erasing improves the accuracy by 0.49% using ResNet-110 while for CIFAR-100 Random Erasing obtains 17.73% error rate which improves 0.76% on the WRN-28-10 baseline.

The research discussed thus far in this section has shown three separate augmentation techniques that have all improved CNN models in some form. However, when implementing augmentation techniques, with so many options to choose from it can be difficult to decide the best augmentation course of action. To help overcome this issue Hussain et al. (2017) performed research looking at different augmentation techniques and the impacts each of them had on CNN results. To compare the results, the different augmentation techniques were applied to a dataset of medical related images, specifically mass and non-mass mammogram images. The researchers applied eight different augmentation types on the dataset and trained eight VGG-16 nets independently on the eight uniquely augmented sets. The specific augmentation techniques used include image mirroring (both horizontal and vertical), gaussian noise, jittering, scaling, powers, gaussian blur, rotations, and shears. Overall the work completed showed that the augmentation strategy, for the most part, greatly affects discriminative performance of CNNs but also drastically affects generative performance, suggesting a strong link between discriminative and generative learning. Of the augmentation strategies applied, image mirroring and gaussian filters performed well leading to validation accuracies of 84% and 88% respectively while strategies such as adding noise to the images actually hindered the performance of the CNN and resulted in a validation accuracy of 66%.

Data augmentation has shown to be a popular method for improving CNN performance when performing classification on images of humans, as seen in the work completed by Hussain et al. (2017), and classification of everyday objects, as seen in work completed by Zhong et al. (2020). This next research paper completed by Wigington et al. (2017) attempts to use data augmentation to improve text classification, specifically recognition of handwritten words. This research focused on the use of a CNN-LSTM model to perform its text classification task. Applying a novel profile normalization technique to both word and line images and augmenting existing text images using random perturbations on a regular grid, the researchers were able to significantly reduce the Word Error Rate and Character Error Rate previously seen in prior handwriting recognition research. The profile normalization technique was applied to reduce the issue seen in the dataset in that differences in handwriting sizes resulted in different image sizes. This technique normalized the data ensuring all images were the same size. From an augmentation perspective, rather than rotating the images of text to increase the data size, the researchers instead employed a random grid mesh to increase or reduce the size of random characters in text. This technique was applied rather than rotating the image as it was identified that rotating the text image was not a natural means of augmenting text data due to that fact that naturally occurring variations in handwriting are not usually manifested as uniform slants across the entire word. The use of the the normalization and augmentation techniques have helped produce the lowest word error rate seen on a number of different text documents, written in different languages and by different authors. Due to the fact that these techniques are not built specifically for the CNN-LSTM model created in this research, an opportunity exists for them to be applied to other hand writing recognition tasks in the future.

As discussed above, data augmentation can improve CNN performance when used in text classification tasks. Further proof that this is the case can be seen in this next research paper completed by Wei & Zou (2019) in which a combination of different augmentation techniques were applied in an attempt to improve CNN performance. The specific techniques applied were synonym replacement, random insertion, random swap, and random deletion. Synonym replacement involves removing a pre-defined number of random words from a sentence that are not considered stop words and replacing them with their synonyms also chosen at random. Random insertion on the other hand is similar to synonym replacement however rather than replacing a word with a synonym, it instead takes a word from a sentence, finds a synonym of the word and inserts it somewhere random in the sentence. Random swap as a technique simply involves choosing two words at random in a sentence and swapping their positions. And finally, random deletion involves deleting each word in a sentence with a pre-determined probability p value. The techniques themselves were applied in

5 separate classification tasks using a mix of full and partial datasets with the partial datasets containing 500 samples. Overall, improvements were seen in the models after the augmentation techniques were applied with a validation accuracy improvement of 0.8% seen in models trained on full datasets and 3.0% seen in models trained on the partial datasets.

Having researched and discussed above a number of examples of the benefits data augmentation has shown to have on CNN performance, it has helped shaped the structure of this research paper and inspire the data augmentation techniques chosen in this research. The results achieved by the research papers above have given more than enough justification for the use of data augmentation in this research, however it is also important to note that not all augmentation guarantees superior CNN performance. The goal of augmenting the data in this research is to attempt to see similar results as seen in the papers discussed above while also gaining an understanding that alternative approaches could yield even better results.

2.1.5 Gaps in Research

The concept of classifying human age and gender from images has not been sufficiently completed, especially when looking at work completed using CNNs. A key gap seen in all previous work revolves around the idea of data augmentation. A small number of studies implemented data augmentation techniques through the use of horizontal image mirroring however more can be done in order to create an overall more balanced dataset. Image rotation will also be used as an augmentation technique to aid in increasing the dataset size. Age imbalance in the dataset was a common theme found in other research in this area. All datasets used had a larger number of images where the subjects were between the ages of 25-40 relative to other age groups. This issue was especially seen with the Adience dataset in which the 25-32 age group contained anywhere from double all the way up to six times the number of photos of any other age group. A technique that was used to resolves this, without reducing the overall number of images in the dataset, was to augment the data using an ageing filter available through the Faceapp application. FaceApp is a popular mobile application for AI photo and video editing. One of the functions of ageing filter applications such as FaceApp is that it allows people in images to be aged older by a significant number of years. Using this application and applying the ageing filter to a set of images will allow the overall dataset to become more balanced in terms of age groups as the number of images of subjects above the age bracket of 25-32 will be increased. Specifically it will be the age bracket of 60-100 that will see an increase in images. This 60-100 age bracket is currently under represented in the Adience dataset. Further to implementing more novel augmentation techniques, more work is required on the comparison of CNN performance versus other model types for age and gender classification. Because of this, a number of SVM models will also be created in this research to help establish if CNNs really are the superior model for image classification.

Chapter 3

Experiment design and methodology

3.1 Design & Methodology

3.1.1 Dataset Description

With the idea of human age and gender classification in mind, one of the first initial decisions to make was to determine the source of data to be used in order to carry out this research. There are a number of datasets in existence that provide the level of information required to perform such a task. Therefore research was conducted in order to determine the most appropriate dataset to use. After much reading and deliberation, a decision was made to go with the Adience dataset. This dataset was chosen as it had been used in multiple other research papers, including research that was for a long time considered the gold standard in this area of study. Due to its use in multiple research studies, this dataset was a natural fit for the age and gender classification tasks this research was aiming to complete.

The Adience dataset was created with human age and gender classification in mind. Further to this, the images included in the dataset were chosen for a specific reason. The creators of this dataset wanted to only include images taken by the average person in everyday life rather than taken by a professional photographer in a professional setting. In other words, the idea was to only include images in everyday, real world situations with natural lighting and poses. The researchers believed this would allow for the efficient creation of models capable of the classifying the age and gender of a human from any image, regardless of the image quality. Unlike other datasets that consist of posed and professional photos, the Adience dataset consists entirely of images of humans that have been taken in everyday life. The researchers who created the dataset deliberately did this as they intended for the data to be as true as possible to the challenges of real-world imaging conditions.

To be particular about what real-world imaging conditions means, it refers to images that contain all varieties of appearance such as variations in angles and lighting, along with different levels of noise and posing. This dataset contains a broad range of angles, lighting, noise, and posing therefore making it a very suitable dataset choice for such a classification task. The images themselves were gathered from multiple albums within Flickr. Flickr is an American image hosting and video hosting service, as well as an online community, similar to that of the more popular Instagram. The initial size of the dataset is 26,580 images of 2,284 unique subjects. The vast majority of images have a corresponding label indicating the gender of the subject in the image along with and the age bracket the subject falls into. There are three distinct gender labels (m, f, u) which represent male, female, and unknown respectively. And there are 8 distinct age brackets (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60-100). In cases where the gender of the subject is 'u' or null these images were removed and in cases where the age bracket was null, these images were also removed. These images were removed as they could not be used in the training process due to their lack of corresponding labels. From initial analysis of the dataset there are slightly more images with an age label than there are with a gender label. The main reason for this being that in a large portion of the images of 0-2 year olds, it is not possible to identify the gender of child and therefore the label is set to 'u'.

The original creators of the Adience dataset made the decision to create a separate version of the dataset. This new version was created in order to aid in the area of age and gender classification. In this version, the faces of the humans in the images were cropped and aligned using a 2D, in plain-alignment tool. In essence, this tool was used to artificially alter the images to try and ensure that the subject of the image was somewhat in-line with the camera and that the focus of the image was the subject's face. The tool was especially useful in cases where the subject of the image may have been off to the side of the image or standing at a distance from where the photo was taken. The alignment tool made it possible to alter these types of images in such a way to allow a model to accurately classify the age and gender of the subject. It was this set of aligned images that was used in this research paper as it was shown in previous research that this version of the dataset resulted in the most successful CNN model (Levi & Hassncer, 2015).

Regarding this research paper, one of the main differentiating factors when compared to previous research in this area was the use of data augmentation techniques. These techniques were used specifically with the idea of creating more data from which the models could train and learn from. The first augmentation technique used was image mirroring which essentially flips the image horizontally in order to create more versions of image from which the models can learn from. Image mirroring was applied to all images in the dataset and was used in both the age and gender classification task. This technique is not particularly novel and has been used in research in the past however it is generally regarded as a beneficial means of augmenting the data. Image mirroring was applied to the dataset as a means of augmentation for a number of reasons. First and foremost, multiple other research papers had cited image mirroring as an excellent means of augmenting the data that often led to performance increases in CNN models. In particular, research performed by Hussain et al. (2017) found image mirroring to be one of the most successful augmentation techniques they tested. And secondly, image mirroring is a cost and computationally efficient means of increasing the data size. In essence, it was a very simple means of doubling the data size for this research.

The second augmentation technique applied on the dataset in this research was the idea of image rotation. Image rotation essentially allows the user to set a specified number of degrees that the image will rotate. By rotating the image, it is essentially shifting the pixels in the image which in turn makes the image appear as a new one to the CNN model when training. Once again, this technique was applied to the dataset for both age and gender classification. The images themselves were each rotated 10 degrees. 10 degrees was chosen as the rotation value as having researched the technique it was discovered that for a task such as human image classification, it was important to rotate the image enough that it appeared to be a completely new image to the CNN but not so much that the image of the person was no longer upright. Having tested multiple images from the dataset, 10 degrees appeared to be an appropriate amount to rotate the images by. Different CNN models were tested with the images rotated using different values. Any rotation degree under 10 resulted in models achieving very similar results however once the images were rotated up to 20 degrees the models tended to disimprove. Similarly to the image mirroring, this is not exactly a new means of augmenting the data and has been used often in the past. However, as mentioned in the literature review, it has helped to boost the performance of CNNs. To further increase the data size once more, both image mirroring and image rotation were used in conjunction with one another. In other words, a new set of images was created by first mirroring the original images and then rotating the now mirrored images. This was a simple and effective means of increasing the data size further with the aim of getting the benefit of two augmentation techniques in one step.

The final and more novel augmentation technique used in this research was the use of a mobile application called FaceApp. FaceApp is an application that can be used to alter the appearance of humans in images. Functionality included in the application includes filters such as swapping hairstyles, adding facial accessories and facial hair, as well as ageing filters. This application was used specifically to augment the data from an age point of view. A flaw or drawback seen and stated in previous research was that the Adience dataset was unbalanced in the sense that there were more images of humans between the ages of 25-32 than any other age groups. This was particularly true when compared to the 60+ age group in which there were far fewer images. This imbalance in the dataset created a bias that resulted in models from previous research being more likely to incorrectly classify humans in the age bracket of 25-32.

Therefore, the purpose of this FaceApp augmentation was to take images of subjects between the age bracket of 25-32 and applying an ageing filter on the images using the FaceApp application. This ageing filter can be used to make subjects of images appear significantly older than they currently are through means such as the addition of wrinkles and greying/whitening of hair. By completing this task, the dataset was increased significantly when it came to subjects in the 60+ bracket resulting in an overall more balanced dataset.

3.1.2 Model Architectures

While the primary focus of this research is to analyse the performance of a CNN at completing an age and gender classification task, a comparison was also made against the performance of Support Vector Machines for both the age and gender classification tasks. Therefore these sections will discuss the process of creating both models and the differences between the two.

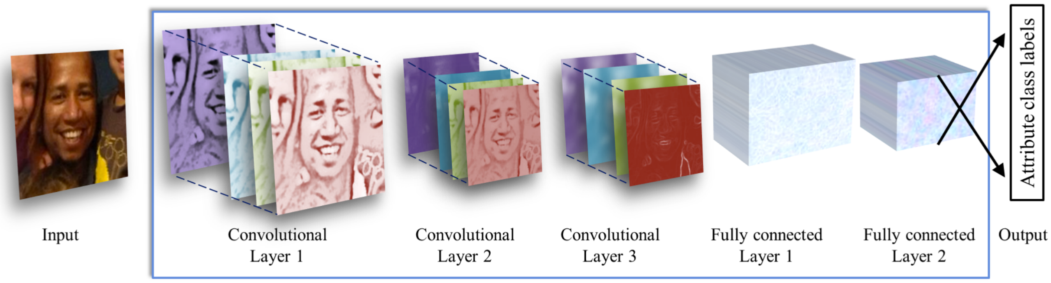


Figure 3.1: Convolutional Neural Network Architecture (Levi & Hassncer, 2015)

CNN Architecture

Five separate CNN models were developed as part of this research, two for performing gender classification (one trained on the standard Adience dataset and one trained on the augmented Adience dataset that included the mirrored and rotated images) and three for performing age classification. The first age classifying CNN was trained on the standard Adience dataset, the second on the augmented Adience dataset that included the mirrored and rotated images, and the third that was trained on the dataset that contained mirrored and rotated images, with the addition of the Faceapp augmented images. Due to the difference in nature between the gender and age tasks, one being a binary classification task and the other being a multi-class classification task, a difference exists between the architectures deployed for each task. Despite the differences in architectures an effort was made to ensure they aligned for the most part. Both models are generally quite simple in nature with a small number of convolutional layers used. This decision was made due to the fact that the classification tasks themselves are not massively complex with only 2 classes in the gender classification task and 8 classes in the age classification task.

This section will speak to the CNN architecture mainly from the gender classification task point of view. For the most part, both the models for each classification task were relatively the same however there are some small features that differ in each. Looking at the gender classifying CNN, the model itself consists of three convolutional layers and two fully connected layers. This architecture choice was made after many rounds of testing in which different numbers of convolutional layers were implemented. Models with everything from one up to four convolutional layers were created and trained on the datasets and it was the three layer model that performed best from both a test accuracy and loss point of view. Another benefit to this architecture choice was that it allowed for a relatively direct comparison to the CNN models created in research by Levi & Hassncer (2015) which is often considered the gold standard of age and gender classification in which the Adience dataset is used. Unlike this research however, rather than using RGB images a decision was made to convert the images to greyscale. The reason for this is that research completed on a similar task found that grayscale images returned better results than RGB images (Ng et al., 2014). Having done some initial testing on a small scale between RGB and grayscale images for this classification task, the conclusion made by Ng et al. (2014) was also true here with grayscale images outperforming RGB. Also colour does not particularly play a role in whether a person is male or female or the age bracket the person falls into so it was assumed the use of colour would not provide a particular benefit. The images were converted to an array and were then normalized by dividing by 255.0 before being passed into a sequential model. Data normalization is an important step in any image processing task as it ensures that each pixel in each image has a similar data distribution. This in turn makes convergence faster when training the network. By dividing the image array by 255.0 it converts the pixels of the images between the values of 0 and 1 which is a suitable range as we require the pixel values to be positive [1]. The structure of the convolutional layers are as follows.

- The first convolutional layer consists of 256 filters with a stride size of 3 x 3. This was directly followed by a rectified linear unit (ReLu) activation function and then a max pooling layer with a pool size 2 x 2 and stride size also of 2 x 2. ReLu is a linear function that will output the input directly if it is positive, otherwise, it will output zero. The input shape is also included here and is the shape of our image data.

- The second convolutional layer and proceeding activation function and max pooling layer are very similar to that of the first convolutional layer. The convolutional layers also consists of 256 filters with a stride size of 3 x 3. The activation function for this layer is ReLu and the max pooling layer has has a pool size of 2 x 2 and stride size of 2 x 2.

- The final third layer convolutional layer once again consists of 256 filters with a stride size of 3 x 3. A ReLu activation function was added once more and the same max pooling layer also followed.

The process of deciding on the hyper parameter values for the convolutional layers involved testing with multiple different CNN models. Having determined that three layers resulted in the best model, it was time to decide on filter size. How this was conducted was by first finding a filter size that resulted in a model that performed relatively well. The initial filter size chosen in this case was 128. Having seen that 128 filters worked decently, a decision was then made to check how 64 and 256 would work

[1] Image Data Pre-Processing for Neural Networks — by Nikhil B — Becoming Human: Artificial Intelligence Magazine, 2017

as filter sizes. A disimprovement was seen with 64 filters while an improvement was seen with 256. Using 256 filters resulted in a model with a higher validation accuracy and lower validation loss. Having seen an improvement by increasing the filter size, an additional check to see how a model with 512 filters would perform was completed. In this instance, a similar validation accuracy was achieved however the validation loss suffered in this instance. It appeared as though some form of over-fitting was occurring with the 512 filter model. At this point a decision was made that 256 filters would work best. This exact same process was followed for the following two convolutional layers with 256 filters being the superior size each time. A 2 x 2 stride size was also determine by comparing it against stride sizes of 1 x 1 and 3 x 3. In this instance, 2 x 2 resulted in the superior output and was chosen as the stride size to continue with.

Following the convolutional layers, the output is then flattened to convert the 3D feature maps to a single long 1D feature vectors. These 1D feature vectors are then passed on to two fully connected layers, once again put in place in an attempt to stay as true as possible to the research completed by Levi & Hassncer (2015).

- The first fully connected layer that receives the flattened input contains 512 neurons and uses a ReLu activation function. A dropout layer of 0.5 was also included here as a means to reduce overfitting. Dropout is used to ignore neurons in the training process of a neural network. This means that their contribution to the rest of the network is temporarily removed.

- The following fully connected layer contains 512 neurons and also uses a ReLu activation function. A dropout layer of 0.5 was also included here as a means to reduce overfitting.

A very similar approach was taken when deciding the hyper parameter values of the fully connected layers as had been done with the convolutional layers. However a number of different models were created before a decision was made to settle with 512 neurons. Initially the first fully connected layer contained 64 neurons and the second contained 32. With these values, the models were performing decently but were not at the level of the models created by Levi & Hassncer (2015). As a result, different neuron amounts were tested going from 32 and 64 up to 126, 256, and finally 512. For both the age and gender classification models it was this 512 value that helped create the best performing models. A trade-off for setting the number of neurons to 512 in both instances was that it took the models longer to train. However the improvements shown in accuracy and loss when testing the different neurons values justified the longer training time. At each of the fully connected layers a dropout value of 0.5 was applied. Dropout is a technique where neurons in the model are selected at random and ignored during training. This means that their contribution to the activation of downstream neurons is temporally removed [2]. Originally the dropout value on both fully connected layers was set to 0.2. During testing this was increased to 0.3 and 0.4 before finally set to 0.5 which resulted in the best model performance. After the data is passed through these fully connect layers the output is moved on to a final dense layer with a sigmoid activation function. It is here that the model determines whether the image passed through is an image of a male of female. Sigmoid is the activation function of choice as this is a binary classification model. This activation function makes use of the logistic regression classification algorithm resulting in output values in the range of 0 to 1. The loss function used for this model is binary crossentropy with the adam optimizer also applied. Binary crossentropy is the loss function of choice as gender classification is a binary classification problem. Binary cross entropy compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalizes the probabilities based on the distance from the expected value [3]. The adam optimizer was chosen as it is often considered the best among adaptive optimizers in most cases. For the sake of exploring more options, other optimizers such as RMSProp and SGM were also tested however it was the adam optimizer that returned the best results.

The second CNN which is used for the age classification task is created in a very similar manner to that of the gender classification CNN. The key differences between the two are that because the age classification task is a multi-class classification task,

[2] Dropout Regularization in Deep Learning Models With Keras, 2016

[3] Binary Cross Entropy/Log Loss for Binary Classification, 2021

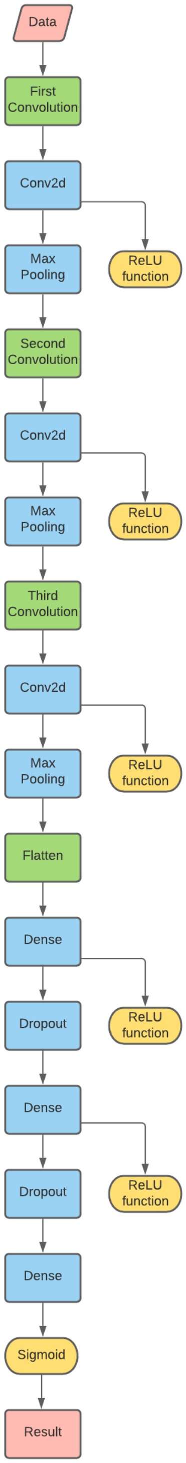


Figure 3.2: Convolutional Neural Network Data Flow - Gender Classification task

the sigmoid activation function in the final layer of the first CNN needs to be changed to a softmax function. As well as that, the number of neurons in the last dense layer needs to be changed from 1 to 8. The requirement for this change is due to the fact that there are eight age groupings from which the model needs to place the subject of an image into. Therefore each neuron represents one of the age groupings. The loss for the model also needs to change from binary crossentropy to categorical crossentropy as this is now a multiclass classification task as opposed to a binary one. The same adam optimizer is used along with the same metrics measured.

SVM Architecture

Support Vector Machines have often been used in the past for image classification tasks, both binary and multi-class classification. Typically SVMs when used for classification purposes are used for binary classification and in doing so try to find a line that maximises the separation between a two-class dataset. In other words the SVM is trying to find the hyperplane that maximises the separation between the two classes. This fact makes SVMs a very suitable model alternative to the CNN for the gender classification problem. In their simplest form, SVMs are not necessarily built to perform multi-class classification. To overcome this issue, SVMs can be used to break down multi-class classification problems into multiple binary classification problems. The two ways to do this are called the one-vs-one approach which breaks down the multiclass problem into multiple binary classification problems and the second approach is the one-vs-rest approach in which the breakdown is set to a binary classifier per each class.

In terms of the SVMs created as part of this analysis because of the difference in the nature between the two tasks, binary classification vs multi-class classification, each of the SVMs implemented in this research were created differently to one another. For the first gender classification task, a binary classification task, the SVM was created with a number of different kernels for comparison purposes. In particular linear, polynomial, and RBF kernels were used in this classification task with the best outputting SVM being used in the comparison between the CNN and SVM. In this scenario the polynomial kernel was the most successful and was the kernel implemented in the final model. The next parameter that needed to be determined was the C value. C is the penalty parameter of the error term. It controls the trade off between smooth decision boundary and classifying the training points correctly. Various values for this parameter were tested before making a final decision. The values tested were 0.1, 1, 10, 100, and 1000 with the C value eventually being set to 1 for the final model [4]. Gamma was the final parameter here that needed to be set. The testing was done between using auto and scale for the gamma parameter before settling on auto for the final model.

For the second age classification task, a further decision was required in the SVM model build due to the nature of multi-class classification. Through researching and applying both the one-vs-one and one-vs-rest approach, the decision made was to go with the one-vs-rest approach. This was due to the fact that with 8 age classes, the one-vs-one approach would have resulted in significantly more models created which was not a realistic solution for this research. Also when attempting to apply the one-vs-one approach, the length of time to train the model was substantially longer than the one-vs-rest. So much so that the time limit for running a notebook cell in Colab was breached and the notebook itself cancelled the running of the cell and reset itself. The final one-vs-rest approach model took approximately 4 hours and 30 minutes to train and numerous articles have stated the one-vs-rest approach is the superior choice of multi-class classification SVM model. ”The obvious approach is to use a one-versus-the-rest approach (also called one-vs-all), in which we train C binary classifiers, fc(x), where the data from class c is treated as positive, and the data from all the other classes is treated as negative” [5].

3.1.3 Training and Testing

Data Preprocessing

The development work of this assignment was completed using Google Colab as the development environment with a number of different notebooks for each of the models created. A GPU was selected for the hardware accelerator as a means of improving on

[4] In Depth: Parameter Tuning for SVC — by Mohtadi Ben Fraj — All Things AI — Medium, 2018

[5] One-vs-Rest and One-vs-One for Multi-Class Classification, 2020

training speed and testing. The notebooks themselves were mounted to a Google drive account where the Adience dataset images were stored. The images were split into different folders corresponding to the different genders and age groups of the image subjects. At this point, externally to Colab, Faceapp had been used to augment the dataset by applying an ageing filter to approximately 500 images from the Adience dataset. These images were then included in the appropriate folders stored on the Google drive account. Having mounted the Colab notebook to the Google drive account, the images were then imported into Colab where some initial preprocessing took place in the form of converting the images from RGB to grayscale. As previously mentioned, research on the use of RGB images versus grayscale has found that grayscale images can perform better for tasks such as gender classification. At this point, with the images now in grayscale, further augmentation techniques were applied. 3 separate functions were created to complete the augmentation. The first function was created to mirror the images, the second to rotate the images by 10 degrees, and the third to first mirror the image and then rotate it 10 degrees. At this point, the dataset was now 4 times the size it originally was.

Having applied each of the augmentation techniques, all the images required to create the different CNN models had been created. However, as each of the images were of different size it was crucial for them to be re-sized to form a consistent size for each of the images in the dataset. A consideration that needed to be taken into account before re-sizing the images was the impact it could have on the quality of the images. To determine an appropriate size to set the images, a large number of different photos were analysed, being set to different sizes. This step was important as it was crucial not to make the images too small that the subject of image could not be seen. After rigorous testing, a decision was made to re-size the images to 100 x 100. Having analysed different image sizes, this 100 x 100 value presented an appropriate balance of reducing the image size to allow for efficient model training while still having images that could be recognised. Recognised, in this sense, means that the gender and approximate age of the subject in the image could be determined by the human eye. A list was then created in which each image was passed to. As each image was passed to the list it was re-sized to our 100 x 100 size choice and the appropriate label indicating the gender or age of the subject was appended to the image. At this point a list of all the images required for training had been created. The images were then shuffled as they had been imported from a Google drive in order of class, meaning that each class was bunched together in the list. Therefore, without shuffling, the images would have been passed one classification after another through the training model which would have severely hampered the models performance. This is because the model would have been trained on specific classes and validation testing would have been completed on different classes leading to poor results. By shuffling the data it ensured the validation or test data were not just one particular class while the training data was another. An error made originally in this analysis was that shuffling had not been put in place and the results of the models reflected that.

Model Implementation

An important feature of this particular analysis is that the models used were created completely from scratch. No form of transfer learning was used throughout the process and only the images from the Adience dataset were used. These are important factors to remember when comparing against previous research in this area in which vastly larger datasets were used (Sun et al., 2014) as it is important to show that complicated larger datasets with multiple features are not always required to complete such a classification task to a successful degree.

Further to what was discussed in the CNN architecture section, a further few methods were implemented in an attempt to improve the results of the models. The use of dropout was implemented as a means of trying to reduce the possibility of overfitting. Dropout has been seen as an effective means of reducing overfitting in research conducted by Park & Kwak (2017) - ”In this paper, we analyze the effect of dropout in the convolutional layers, which is indeed proved as a powerful generalization method. We observed that dropout in CNNs regularizes the networks by adding noise to the output feature maps of each layer, yielding robustness to variations of images.”

The models were first created using the dataset without any augmentation techniques implemented. The results were analysed and stored for future comparison with the models in which data augmentation was applied. The augmentation techniques as previously discussed were the use of image mirroring, image rotation, and also the use of the FaceApp application filters to make the subject in the image appear older, in essence creating new images of older subjects. The models were ran using 100 epochs and a batch size of 32 with a training/validation/test split of 80/10/10. 100 epochs were chosen to give the models an opportunity to train for as long a time as possible, as long as the model results were continuously improving. Given the amount of training samples, a batch size of 32 was selected. 64 and 128 were alternative batch sizes that were tested however 32 appeared to be the optimum solution. Regarding the training/test split, 80/20 often tends to be a go to split in the deep learning space. However with the extra validation metrics included a decision was made to provide 10% of the data for validation purposes and to keep 10% of the data unseen to the model for testing purposes. Early stopping techniques were applied to the models to prevent the models' performance decreasing throughout the training process. Early stopping is used to stop a model's training process if the results of the model begin to disimprove before the number of training epochs has been completed. In these instances early stopping was implemented with respect to validation loss. A patience of 3 was put in place meaning that the models would continue to train up until the point that the validation loss did not improve within 3 epochs or the 100 epochs of training had surpassed. A patience of 3 was chosen after a number of different patience values were tested. 3 was chosen as the final value as higher patience values, for example 5 or 10, tended to result in models where the validation loss would have diminished significantly away from the minimum validation loss achieved during training.

With respect to the metrics that were included in the analysis of the model, as standard, the accuracy and loss of the models were the first two metrics included. Accuracy is a measure of how accurate a model's predictions/classifications are compared to the true data value. Loss on the other hand is the sum of errors made for each example in training or validation sets. Loss value implies how poorly or well a model behaves after each iteration of optimization. These two metric types are often the metrics quoted in other research papers however other metrics such as precision and recall can also be taken into account if necessary. Precision and recall are metrics considered in this research paper as a means of further understanding model performance. In particular, precision and recall were analysed to help understand individual class classification performance.

Hypothesis

The original state-of-the-art model for classifying the age and gender of humans from images was created by Levi & Hassncer (2015). This research focused on creating a convolutional neural network using three convolutional layers and two fully connected layers. Using the Adience dataset, which consists of 26,580 images of 2,284 subjects, an accuracy of 86.8% for classifying gender and 50.7% for classifying age was achieved.

H_0 - The state-of-the-art model for classifying the age and gender of people from images of humans uses a convolutional neural network comprised of three convolutional layers and two fully connected layers and results in an accuracy of 86.8% for classifying gender and 50.7% for classifying age.

H_1 - Through the use of data augmentation techniques, specifically horizontal image mirroring and image rotating, an accuracy of greater than 86.8% for classifying gender can be achieved.

H_2 - Through the use of data augmentation techniques, specifically horizontal image mirroring, image rotating, and Faceapp image filtering, an accuracy of greater than 50.7% for classifying age can be achieved.

The objective of this research is to prove that while the results of previous research in this area achieved excellent results, improvements to these results can be made with data augmentation techniques. The specific augmentation techniques include the use of image mirroring, image rotation, and also the application of an age filter to the images. Horizontal image mirroring as well as image rotation will be applied to boost the number of images in the dataset. The FaceApp application will also be integrated into the research. This application allows for ”filters” to be applied to images of humans. These filters allow for the increase/reduction in age of the subject in an image by a particular number of years. This augmentation technique will aid in diversifying the age groups of the subjects in the images. In particular, due to the nature of the fact that there are far few images of subjects over the age of 60 in the Adience dataset, the ageing filter will be applied to a number of images of subjects between the ages of 25-32 to help boost the number of images of this age group. This will be done with the intention of improving the balance of the age groups in the dataset.

The second key objective of this research was to establish which model type, between a CNN and an SVM, is the superior for performing age and gender classification using the Adience dataset. CNNs have widely been considered the go to model type for completing image classification tasks. After gaining interest from researchers in the 1990s, CNNs really surged in popularity in the 2010s after a CNN model AlexNet managed to achieve state-of-the-art results in an image classification task called the ImageNet challenge. From this point, CNNs have been the model type of choice for researchers looking to complete image classification tasks. However before CNNs reached the height of their popularity, other model types were required to perform these image classification tasks with SVMs being a popular choice in this regard. SVMs were first created back in the 1960s before being refined in the 1990s. SVMs are a supervised classification algorithm which are more than capable of classifying images. Built naturally for binary classification tasks, an SVM model can easily be applied to the gender classification task being analysed in this research paper. However thanks to the development of methods that can be applied to SVMs such as the one-vs-one and one-vs-rest methods, it is also possible to apply an SVM to the age classification tasks being analysed here. The goal here will be to apply CNNs and SVMs on the same Adience dataset using the same conditions, to determine which of the models perform the tasks to a higher level.

Due to the nature of the fact that this research is an extension and an attempt at improvement over the research completed by Levi & Hassncer (2015), it is that research from which we will compare results. As both sets of research made use of the Adience benchmark dataset and implemented similar model types in the form of CNNs, this allowed for a direct comparison from which to be made. Because of this direct comparison, it was possible to determine whether CNNs are superior to SVMs and also whether the data augmentation techniques implemented did have a beneficial result for the models created. This would allow for the accepting of the alternate hypothesis and accepting of the fact that a more balanced dataset in terms of age would result in overall better model performance.

Chapter 4

Results, evaluation and discussion

4.1 Results

In this section the results of the various models created throughout this research process will be laid out. The results were generated using a number of CNN and SVM models with varying degrees of data augmentation applied. In relation to the tasks of classifying gender and age, the initial models created were a CNN and SVM with just the bare minimum dataset used in the training process. In other words, the models were created with none of the previously discussed data augmentation techniques applied. These models were created as a form of baseline from which comparisons could be made when new models with the data augmentation techniques were implemented. Having created the baseline models, data augmentation techniques were applied to the dataset to determine whether they were beneficial or unfavourable to the models in terms of performance. For both the gender and age classification tasks, the specific augmentation techniques applied were to mirror the images horizontally in order to create ”flipped” versions of the images. By doing this it essentially doubles the overall size of the dataset with each image being included in the mirroring process. The next technique applied was to rotate the original images a specific number of degrees. This is another common data augmentation technique as it is a simple process that allows the user to select exactly the number of degrees the images should be rotated. For this research, the images were rotated by 10 degrees. 10 degrees was chosen after experimenting with a number of different values as it ensured that the images were still relatively upright for the model to train effectively. The 10 degree value also allowed for enough variation from the original images that the model would not essentially be being trained on the same images again. Finally, a combination of both the horizontal image mirroring and image degree rotation techniques were applied to the images. Having applied each of the augmentation techniques, the resulting dataset was now 4 times the size of the original. These augmentation techniques were implemented as a result of research completed by Shijie et al. (2017). This research, which looked at the impact of various augmentation techniques on the results of convolutional neural networks, found that image mirroring and rotating were two of the more successful techniques with classification performance improving by between 3% and 3.5% on the CIFAR-10 dataset.

The next data augmentation technique was applied to the age classification task only. This is because the purpose of it was to solve an issue seen in previous research papers, such as the paper completed by Levi & Hassncer (2015), in that the age classification results were significantly lower than that of the gender classification, in part, due to the fact that there is an imbalance in the dataset from an age bracket perspective. In the Adience dataset, there are more than twice the number of images for subjects between the ages of 25-32 than any other age bracket. This resulted in models that were more likely to incorrectly classify subjects in this 25-32 age bracket. Therefore, the Faceapp mobile application was used as a means of augmenting the data from an age perspective through the use of filters. Filters included in the application include new hairstyles, facial hair, and accessories such as sunglasses. But most importantly it allows for an ageing filter to be placed on a person. This ageing filter can be used to make a person appear younger or older than they currently are. For this research, the filter would be applied to make the people in the images appear older. The reason for this is that there is a significantly smaller number of elderly subjects (60 years and above) in the dataset than any other age group. The ageing filter makes people appear older through a number of means such as increasing the number of wrinkles the subject has while also often changing the subjects hair colour to grey/white as would naturally tend to happen in old age. Applying this filter would allow for a more balanced dataset with the goal of reducing the issue faced in other research papers that found subjects being incorrectly classified in the 25-32 age bracket. A potential issue with the introduction of this filter is the issue that the application assumes that all people age in the same manner with the same ageing attributes such as wrinkles and hair colour change. While this is true in the majority of cases, there are exceptions out there in which these ageing attributes do not apply. This is a relatively small issue to worry about but it is worth pointing it out regardless.

Below is a breakdown of the number of images per classification in the dataset before and after data augmentation has occurred.

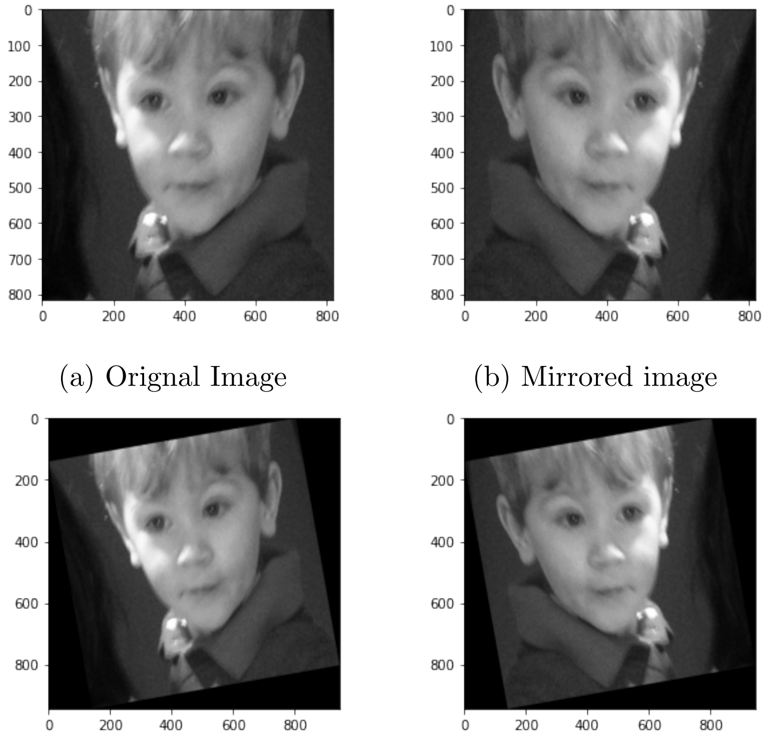
	Pre Augmentation	Post Augmentation
Male	8,090	32,360
Female	9,343	37,372
Total	17,433	69,732

Table 4.1: Dataset breakdown by gender - pre/post augmentation

	Pre Augmentation	Post Augmentation
0-2	2,487	9,948
4-6	2,140	8,560
8-12	2,125	8,500
15-20	1,641	6,564
25-32	5,102	20,408
38-43	2,346	9,384
48-53	830	3,320
60-100	873	5,424
Total	17,544	72,108

Table 4.2: Dataset breakdown by age - pre/post augmentation

For the most part, in Tables 4.1 and 4.2, the post augmentation value is simply 4 times the pre augmentation value. The only exception to this is in the 60-100 age bracket where the Faceapp augmentation has also occurred. The original number of subjects in the 60-100 age bracket was 873. After the Faceapp augmentation this increased to 1,356 before finally reaching 5,424 after the mirroring and rotating augmentation were also applied.



(a) Orignal Image

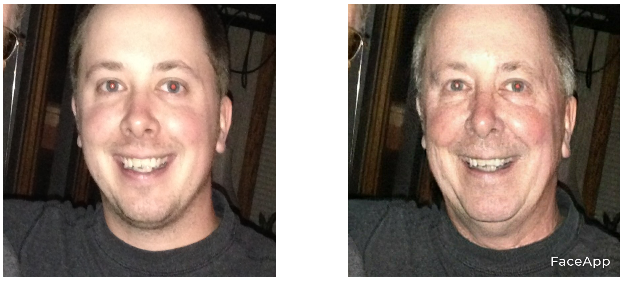
(b) Mirrored image

(c) Rotated Image

(d) Mirrored & Rotated Image

Figure 4.1: Example of Image Mirroring & Image Rotation

The below tables, 4.3 and 4.4, contain the results of the models created for both the age and gender classification tasks. The results quoted in the table are both the pre and post augmentation results. Unfortunately, in the case of the SVM model, it was not possible to apply the post augmented dataset in the same manner as the CNNs. The reason for this is due to the time limit applied by Google Colab when running a notebook cell. The training time of the SVM for the augmented dataset breached the time limit of 12 hours applied by Google Colab for a notebook cell runtime. A



(a) Image without Faceapp filter

(b) Image with Faceapp filter

Figure 4.2: Example of image pre and post Faceapp filter

workaround for this was to split the augmented dataset into 3 separate datasets and to train a different SVM on each set. Having completed this, the average accuracy of the 3 SVM models was taken as the final result. The reason for splitting the dataset into 3 parts was to condense the data to allow an SVM to train within the Google Colab time restrictions while also allowing for the SVM models to be trained on some form of augmented dataset. As the original dataset was augmented to 4 times its original size, by splitting the augmented data in 3 it still allowed for each SVM to train on a larger dataset than the original.

Gender Classification	Pre Augmen. Accuracy	Post Augmen. Accuracy
CNN	86.9%	89.5%
SVM	74.1%	71.5%*

Table 4.3: Gender Classification Results - pre/post augmentation. *Not fully augmented dataset

Below are a number of confusion matrices, Figure 4.3 and 4.4, providing breakdowns of the predicted values made by the CNN models versus the actual labels associated with the images. There are four confusion matrices included to provide a further direct comparison of the results of the CNNs trained on the standard dataset versus the CNNs trained on the augmented dataset. The first set of confusion matrices

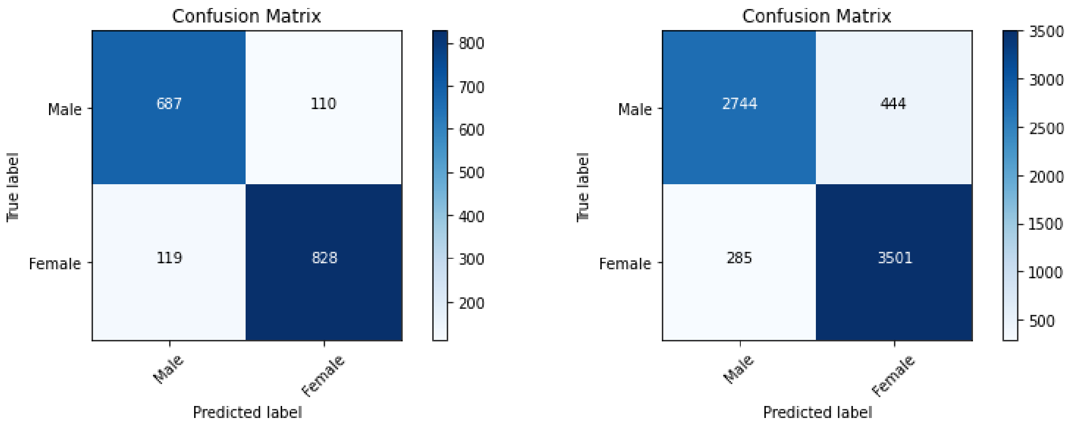
Age Classification	Pre Augmen. Accuracy	Post Augmen. Accuracy
CNN	65.1%	72.2%
SVM	43.9%	37.9%*

Table 4.4: Age Classification Results - pre/post augmentation. *Not fully augmented dataset

provides the classification breakdowns made by the gender classification CNNs while the second set of confusion matrices looks at the classifications made by the age classification CNNs. This age classification CNN will be looked at in determining whether the introduction of the data augmentation techniques helped to reduce the issue of the one-off problem. The one-off problem is an issue seen in age classification where a subject is classified by a model into the age bracket directly above or below the actual age bracket they below to. An example of this would be a subject who is in the age bracket of 25-32 being classified as a 15-20 year or 38-43 year old. A reduction in the one-off problem in these circumstances is simply a model that is capable of classifying the age bracket of a subject with more accuracy. The augmentation techniques, especially the Faceapp augmentation, were applied to the dataset specifically to try to reduce this issue of the one-off problem.

4.2 Discussion

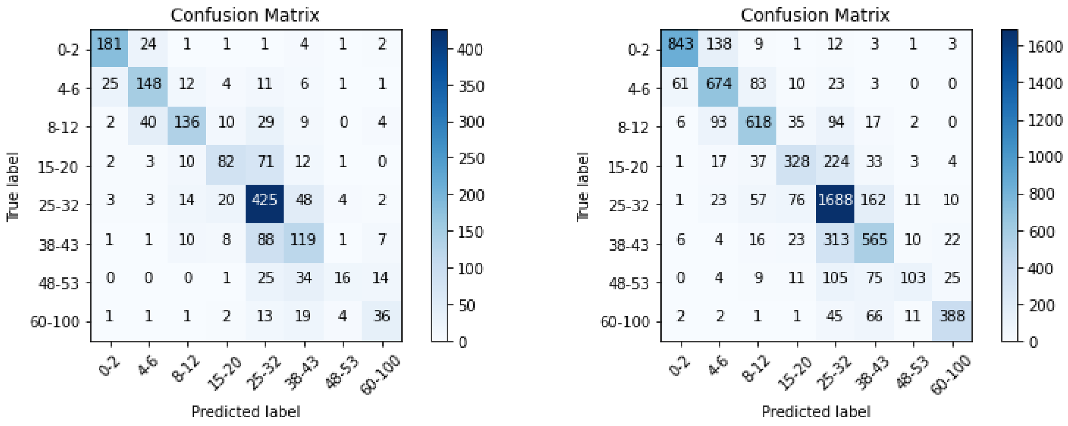
The purpose of this research was to attempt to improve on the previous state of the art work completed by Levi & Hassncer (2015) through the use of both well known and novel data augmentation techniques. As previously discussed, the augmentation techniques include image mirroring, image rotating, and the use of the ageing filter in the Faceapp application. The idea behind employing these techniques was to aim to improve on the accuracy scores of the models created by Levi & Hassncer (2015). Data augmentation techniques have been found to be an effective and cost efficient means of improving CNN accuracy (Halevy et al., 2009).



(a) Gender classifications pre-augmentation

(b) Gender classifications post-augmentation

Figure 4.3: Confusion matrices of gender classification models



(a) Age classifications pre-augmentation

(b) Age classifications post-augmentation

Figure 4.4: Confusion matrices of age classification models

Looking initially at the gender classification results and focusing on the pre-augmentation accuracies achieved by both the CNN and SVM. Part of this research was to compare the effectiveness of CNNs at performing image classification versus other model types that would have been used in the past. Before CNNs became the gold standard for image classification, SVMs would have been a very popular model choice for such a task. It was therefore chosen as the natural model type for which to compare the

CNN results against. Before performing any augmentation techniques to the dataset it is clear to see that the CNN outperformed the SVM with regard to gender classification. The CNN achieved an accuracy result over 12% greater than than of the SVM making it a superior model choice for gender classification with the Adience dataset. In terms of results before and after the augmentation techniques were applied, it can be seen that the augmentation appears to have been the cause of the improvement between the CNN models. For the gender classification task it was just the image mirroring and image rotation augmentation techniques that were applied. The Faceapp augmentation was specifically for the age classification task. After the augmentation had been applied the accuracy of the CNN increased from 86.9% to 89.5%, a 2.6% increase which is significant when the pre-augmentation result was already relatively high. Unfortunately, due to the time-out limitations associated with running notebook cells in Google Colab, it was not possible to train an SVM model with the same fully augmented dataset that the CNN was trained on. Instead, the augmented dataset was split in 3 with an SVM trained on each data split and the average metric results used as the final result. As mentioned previously, by splitting the augmented dataset in 3, each split was still larger than the original dataset. Interestingly, the average result of the 3 augmented SVMs was lower than that of the SVM trained on the original dataset. The average accuracy result of the 3 augmented SVMs was 71.5%, down from 74.1%. None of the accuracy results of the 3 augmented SVMs was higher than 74.1%. This finding could indicate that data augmentation is not always beneficial to SVM models.

When comparing the pre-augmentation CNN gender classification accuracy of this research to the gender classification accuracy achieved by Levi & Hassncer (2015), the results are nearly identical. The CNN of Levi & Hassncer (2015) achieved a gender classification accuracy of 86.8% while the CNN of this research achieved an accuracy score of 86.9%. However with the augmentation techniques applied, a significant performance benefit can be seen in the CNN model created as part of this research. And when compared to the CNN created by Levi & Hassncer (2015), this research's CNN has a 2.7% higher accuracy. Therefore with respect to the gender classification task it is possible to accept the first alternative hypothesis proposed by this research. This is possible as the augmentation techniques have resulted in a model with a gender classification accuracy greater than 86.8%.

Moving on to the age classification results, and in particular the pre-augmentation results, it can be seen that the CNN outperformed the SVM in this task with accuracy scores of 65.1% and 43.9% respectively. Having analysed the gender classification results prior to looking at the age classification task, it was expected that the CNN would once again outperform the SVM however to do so by approximately 21.2% is a significant outperformance. Moving on to the pre and post-augmentation results, once again an improvement in CNN accuracy can be seen once the data has been augmented. For the age classification task the data was augmented using the three techniques discussed in this paper - image mirroring, image rotating, and using Faceapp. With the augmentation techniques applied an improvement in accuracy score from 65.1% to 72.2% was seen. With such a vast improvement in accuracy of this Faceapp augmented model it would appear on the surface that the issue of the one-off problem would have been reduced. The one-off problem is an issue seen in previous age classification research papers where models classify the subject of the image to be in the age bracket directly above or below the actual age bracket they belong to. And while an increase in model accuracy is a great step in the right direction to improving on this issue, further analysis through confusion matrices will be discussed further in this paper to confirm whether the one-off problem has been reduced or not.

Similarly to the SVM used in the gender classification task, the same issue was faced when training the SVM for age classification with the augmented dataset. The runtime of the training was longer than the notebook cell runtime that Google Colab permits so once again the augmented dataset was split in 3 with the average accuracy result of the 3 SVMs taken as the final result. Once again the average result of the 3 SVMs trained on augmented data was lower than that of the SVM trained on the standard dataset. The average accuracy result of the 3 SVMs was 37.9%, down from the 43.9% achieved by the original SVM. This is further evidence the augmenting data does not always result in greater performance for SVM models.

Now once again comparing the results of this research to that of the results of Levi & Hassncer (2015). The age classifying pre-augmented CNN of this research outperformed the CNN of Levi & Hassncer (2015) rather significantly. The CNN of this research achieved 65.1% for age classification while the CNN of Levi & Hassncer (2015) achieved 50.7%. And while the models were designed very similarly to one another, with both having three convolutional layers and two fully connected layers, it is possible that the use of greyscale images in this research was a key factor in the better results. As mentioned previously in this paper, previous research completed by Ng et al. (2014) found that greyscale images for classification tasks can often lead to better results. And with the CNN accuracy of this research improving further post data augmentation, from 65.1% to 72.2%, we can accept the second alternative hypothesis that the augmentation techniques applied in this research can result in a better performing CNN than that one the CNN created by Levi & Hassncer (2015). Overall, this research could be a significant finding for this area of work. The ageing filter functionality such as that provided by the Faceapp application is still relatively in its early stages of development. As this type of technology improves and an ageing filter can be created that ages somebody by an exact number of years, the ability to create new images increases significantly, potentially leaving the issue of a lack of data in the past.

To further expand on the results, it was important to look at confusion matrices of the classified versus true labels. Classification accuracy is a useful metric when looking at the performance of models, however the addition of confusion matrices can help overcome the sometimes misleading nature of classification accuracy, particularly when dealing with multi-class classification problems. Confusion matrices are beneficial for multi-class problems as they help to determine the classification performance of each individual class. For example, a model may have an overall accuracy of 75% however in a multi-class problem it may be a case than one class is predicted correctly 100% of the time, the second class 75% and the third only 50%. Without compiling a confusion matrix in this scenario, this insight would not have been identified. Therefore there is a direct benefit of looking at a confusion matrix for the age classification task as there are eight distinct age groupings within the Adience dataset making it a multi-class classification problem. Additionally confusion matrices are also beneficial when the dataset used in creating the model has some form of imbalance. This is because it is possible for a model to have an accuracy of 90% however if 90% of the data instances all fall into the one class then it is possible that the model just predicted the same class every time to achieve this 90%. This reasoning shows another benefit of looking at a confusion matrix for this research, for both the age and gender tasks. With the age data in the Adience dataset there is an imbalance of instances in the 25-32 age bracket with this bracket containing more than double the number of instances than any other age bracket. And with the gender data there are slightly more images of females in the dataset (53%) than males (47%) so a confusion matrix is also beneficial here despite the imbalance of instances being relatively small.

The confusion matrices included in this paper are the matrices generated specifically from the CNN model outputs. Starting with the gender classification confusion matrices, there were only two genders from which the model was to classify. This section will discuss the results of the matrices from an accuracy, precision, and recall perspective. In it's simplest form, precision is the ratio between the true positives and total positives (true positives plus false positives). A true positive result means that the model correctly predicted the positive classification for the test sample. A false positive result means that the model incorrectly predicted the positive classification for the test sample. Recall on the other hand calculates the number of positive class predictions made out of all positive instances in the dataset. As each of the classes are equally as ”important” as one another, precision and recall will be calculated from both genders point of view. Looking firstly at the confusion matrix created from the output of the CNN trained for gender classification with no augmentation applied to the dataset. This is Figure 4.3(a) above and it can be seen that the accuracy of the model was 86.9% or .869. This is calculated by summing the number of correct classifications made by the model and dividing by the number of overall classification made. This calculation is (1,515/1,744) * 100 = 86.9%. Now while this accuracy metric is useful as an initial indication of model performance, it can also be important to look at the precision and recall of a model's output. Precision is calculated by dividing the number of true positive cases by the sum of true positive plus false positive cases. Looking first at the male classification results, in this instance the number of true positive cases is 687 (the number of correctly classified males instances) and the number of false positives is 119 (the number of females classified as males). These figures give a precision result of .852 (687/(687+119)). Looking next at the precision from the female classification point of view. The number of true positives is 828 while the number of false positives is 110 (the number of males classified as females). This gives us a precision value of .883 (828/(828+110)). Both the male and female precision values are excellent results. The fear in this scenario is that there would be a large difference in the result of the model accuracy versus model precision. Fortunately that is not the outcome in this instance indicating that the model is indeed performing well. A potential reason for the slightly better performing model from a female classification point of view could be the larger amount of female image data available that the model was able to train on.

Moving on to look at the model recall - recall is calculated by dividing the number of true positives by the sum of true positives plus false negatives. A false negative is a case when a model wrongly classifies that a particular label or attribute is present. In the context of male classification in this task, a false negative would be the model predicting a male as female. And from a female classification perspective a false negative would be the model classifying a female as male. Looking first at recall from the male classification view, the number of true positives is 687 while the number of false negatives is 110. This gives the model a recall value of .862 (687/(687+110)) from a male classification point of view. From the perspective of female classification, the number of true positives is 828 while the number of false negatives is 119. Calculating recall using these figures gives a value of .874 (828/(828+119)). Similarly to the precision results achieved by the CNN model it is important that the recall values align somewhat with the model accuracy. In doing so it shows that the model accuracy result is fair and an accurate representation of the actual model performance.

Having discussed above the confusion matrix results from the model that was trained with no augmentation applied to the data, it is now time to look at how that model compared to the CNN trained with the augmented dataset. This confusion matrix can be seen in Figure 4.3(b) above. The accuracy of this overall model was 89.5% or .895 which can be calculated by dividing the sum of the correct classifications by the sum of all the calculations - (6245/6974) * 100 = 89.5%. Once again this result on paper looks excellent however it is important to again look at the precision and recall results of the model. Starting with the male classification precision results, the total number of true positive classifications was 2,744, while the number of false positives is 285. This gives a precision values of .906 (2744/(2744+285)). This is an excellent result achieved by the model with a clear direct improvement on the CNN trained on the standard dataset. This result also lines up with the overall model accuracy improvement seen with this model. Moving next on to the female classification precision results. In this instance the number of true positives was 3,501 with the number of false positives being 444. These values give a precision result of .887 (3501/(3501+444)). Once again this is another excellent result and shows that the model is performing impressively for both the males and female classification task, from the precision point of view at least. An interesting point to note with the precision results of the CNN trained on the standard dataset versus the CNN trained on the augmented set. With the standard trained CNN it was the female precision result that was the slightly higher than the male. However with this augmented trained CNN the opposite can be seen with the male precision score being slightly superior. This means that for the original CNN model, it was more likely to classify a female as male while in the augmented CNN model it was more likely to classify a male as female. In both models, the difference between the male precision score and female precision score is relatively minor but a point of interest nonetheless. From this result it can be said that the augmentation techniques improved the precision result of the male classification more than the female.

The next step in the process was to then look at the recall results of the CNN trained on the augmented dataset. As previously mentioned, recall is calculated by dividing the number of true positives by the sum of true positives plus false negatives.

For the male classification recall results, the number of true positive was once again 2,744. And the number of false negatives was 444. This gives a recall result of .861 (2744/(2744+444)). This is a slightly lower score than the precision achieved by the CNN model however still an impressive score regardless. On the other hand, looking at the female classification recall results, the number of true positives in this case was 3,501 while the number of false negatives was 285. These figures give a recall result of .925 (3501/(3501+285)). This recall result is the highest achieved by any model conducted in this research. Now comparing these recall results against the recall results achieved by the original model. The recall result from the male classification point of view disimproved slightly dropping from .862 to .861. The difference in recall results of the two models is so small that it essentially means they are both equally likely to classify an image of a male as a female. On the flip side with the female classification recall result, an improvement was seen from the original model to the augmented model from .874 to .925. This improvement means that the augmented model is less likely to incorrectly classify an image of a female as a male than the original model.

Having delved further into the results of each gender classification model, looking beyond just the accuracy metric to explore the precision and recall results, it is evident that the gender classification model has seen improvements having been trained on an augmented dataset. One of the biggest goals of this research was to establish if more training data could improve the results of a CNN in a gender classification task. And looking at the increases made in accuracy, precision, and recall from the original CNN to the augmented CNN it is fair to say that improvements have been seen. And while the recall result for the male classification aspect of the CNN did not improve with a slight drop seen from .862 to .861, this does not detract from the overall improvement seen in the model.

The next set of results that will be discussed here are the precision and recall results of the age classification models. There are a number of figures and tables that can be used to analyse these results. Figure 4.4(a) and 4.4(b) are the confusion matrices for the age classification results pre and post data augmentation while Tables

Gender Classification	Pre-Augmentation	Post-Augmentation
Male Precision	.852	.906
Female Precision	.883	.887

Table 4.5: CNN Gender Classification Precision - pre/post augmentation

Gender Classification	Pre-Augmentation	Post-Augmentation
Male Recall	.862	.861
Female Recall	.874	.925

Table 4.6: CNN Gender Classification Recall - pre/post augmentation

4.7 and 4.8 provide the actual precision, recall, and F1-scores of the respective models. F1-score is the weighted average of precision and recall. Starting with the precision and recall scores of the CNN model trained on the standard dataset. From both a precision and recall perspective, the age bracket of 0-2 achieved the highest results of .842 each giving an F1-score of .842 also. Comparing these results to the rest of the age brackets, these .842 scores are vastly superior to the scores of the other age brackets with the next closest age bracket being 25-32 achieving an F1-score of .719. The reason for this age-bracket performing so well is likely down to two main reasons. The first being the fact that after the 25-32 age bracket, this 0-2 bracket contained the highest number of images. And while there were not significantly more images in this 0-2 age bracket than the other age brackets, it is possible that the extra number of images may have made a difference in this case. The second possible reason for this result could be down to the fact that humans between the ages of 0-2 look significantly different to the other age groups in the dataset. For example, there is a large difference in appearance between a 0-2 year old and a 4-6 year old. However the same could not be said to the same extent for the difference between a 38-43 year old and 48-53 year old. The difference in appearance between a 0-2 year old and the next age bracket up is far larger than the difference between any of the other age brackets and their next corresponding bracket. Therefore it is likely that this vast difference in appearance between 0-2 and any other age group allowed for the model to learn this age bracket very well, resulting in the high precision and recall scores for 0-2 year olds. On the flip side, the worst performing age bracket was that of images of 48-53 year olds. This bracket had a respectable precision score of .571 however the recall score was the key reason for the poor performance only reaching .178. This means that although the model would have an average number of false positive results for this age bracket it would have a very high number of false negatives which is a disappointment in this case. The most likely reason for this poor result is simply down to a lack of data. Both the 48-53 and 60-100 age brackets had a very similar number of images to train the model on. However the key difference between these two sets of images is that someone over the age of 60, especially a person who is closer to 80 years old and above, would have a much more distinctive look, especially when compared against the other age groups, than a person between the age of 48-53. Looking at the confusion matrix for this original age classifying CNN, it can be seen that there were more 48-53 year olds classified as 38-43 and even 25-32 year olds than in their correct age bracket. There were 16 48-53 year olds classified correctly while 34 of them were classified as 38-43 and 25 of them were classified as 25-32. It is these incorrect classifications that have resulted in such a low recall score. And as mentioned previously this is likely due to the lack of training data for the model to learn on.

Turning the attention to the confusion matrix associated with the age classifying CNN trained on the augmented dataset now. In this section the relevant precision, recall, and F1-scores will also be discussed. As was established earlier, the accuracy of the augmented model was superior to that of the original CNN. However with this confusion matrix and additional metric scores it is possible to delve further into the performance of the CNN and how it compares to the original. In the original CNN it was the 0-2 age bracket that performed best in terms of precision, recall, and resulting F1-score. This is the same case for the augmented model however the scores between the 0-2 age bracket and the rest are far closer. The 0-2 age bracket was already

Age Bracket	Precision	Recall	F1-Score
0-2	.842	.842	.842
4-6	.673	.712	.692
8-12	.739	.591	.657
15-20	.641	.453	.531
25-32	.641	.819	.719
38-43	.474	.506	.490
48-53	.571	.178	.271
60-100	.545	.468	.503

Table 4.7: CNN Age recall, precision, and f1 scores - Pre Augmentation

performing quite successfully before the augmentation so a major improvement on this was always going to be difficult. However to see a great improvement in the other age brackets is a promising sign. The 0-2 age bracket achieved a precision score of .916 and a recall score of .835 giving a final F1-score of .874. Similarly to the original CNN the 48-53 age bracket was once again the worst performing age bracket from an F1 score perspective. However significant improvement was made post the augmentation. The precision of this age bracket increased from .571 to .730 while the recall score increase from .178 to .310. This gave an overall F1-score increase of .165 (.436 - .271). For this CNN model the 48-53 age bracket once again had the smallest number of training samples however the improvement in scores has shown the benefit that more data had in increasing the performance of the CNN.

The most improved age bracket from the original CNN to augmented CNN was the 60-100 age bracket. This is significant as this was the age bracket that received the most augmentation through the image mirroring, image rotation, and Faceapp augmentation. The 60-100 age bracket was the only bracket to receive this Faceapp augmentation as the ageing filter was used to increase the number of images of subjects in this age bracket. The precision score increased from .545 to .858 while the recall

Age Bracket	Precision	Recall	F1-Score
0-2	.916	.835	.874
4-6	.706	.789	.745
8-12	.745	.714	.729
15-20	.676	.507	.580
25-32	.674	.832	.745
38-43	.611	.589	.600
48-53	.730	.310	.436
60-100	.858	.752	.802

Table 4.8: CNN Age recall, precision, and f1 scores - Post Augmentation

score increased from .468 to .752. This resulted in an overall F1-score increase from .503 to .802. This is an extremely important finding from this research as it shows directly the positive impact increasing the dataset size through Faceapp has had on the model results and allows for the second alternative hypothesis to be accepted. This second alternative hypothesis was that 'Through the use of data augmentation techniques, specifically horizontal image mirroring, image rotating, and Faceapp image filtering, an accuracy of greater than 50.7% for classifying age can be achieved'.

Moving on, one of the key problems in the area of age classification that this research was looking to improve on was the issue known as the one-off problem. As previously mentioned, the one-off problem is an issue in age classification in which the model classifies a subject to be in the age bracket directly one above or below the actual actual age bracket of the subject. The reason the one-off problem occurs is that from one age bracket to the next there is not as much of a difference in appearance as there would be in a two bracket or more age difference. For example, a 25-32 year old will typically look more similar to a 38-43 year old than they will a 48-53 year old. This is an issue seen especially in the research completed by Levi & Hassncer (2015) in which their model achieved an age classification accuracy of 50.7% for an exact age bracket classification and an accuracy of 84.7% for a one-off age bracket classification. As has been seen with this research the best performing augmented CNN has already performed better than the best performing CNN of Levi & Hassncer (2015). This research's model achieved an age classification accuracy of 72.2% compared to the 50.7% of Levi & Hassncer (2015). This improvement in accuracy result has already indicated a reduction in the one-off problem itself. Also the improvement in precision and recall values for each of the age brackets is another a key indication of a reduction in the one-off problem. Increasing precision and recall results indicate that the number of false positive and false negative classifications produced by the model have both decreased resulting in a reduced one-off problem.

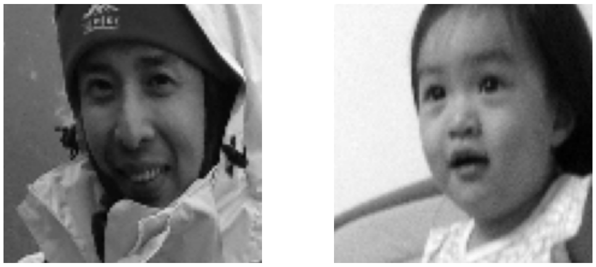
And while a key part of this analysis was to increase the accuracy of the CNN model to reduce the one off problem, it is also important in this analysis to looking at the model accuracy from a one-off classification point of view. In other words, it's important to look at the accuracy of the model at classifying a subject to the exact age bracket or the age bracket directly one above or below their actual age bracket. This metric holds significance as it is an extra result produced by the model of this research that can be compared against the research of Levi & Hassncer (2015). As well as this, it also helps to show whether or not the model is classifying the subjects in the dataset close to the age bracket they actually fall in to. For example with the model created by Levi & Hassncer (2015) that had a one-off classification accuracy of 84.7%, this means that the model misclassified subjects by at least 2 age brackets or more 15.3% of the time. It was therefore important to try to increase this one-off classification accuracy in the models created in this research. Fortunately an improvement was seen with this research as the augmented CNN achieved a one-off classification accuracy of 90.8%. This means that only 9.2% of classifications made by the model were 2 or more age brackets away from the correct age bracket of the image subject, a 6.1% improvement on the model created by Levi & Hassncer (2015).

While the alternative hypotheses have been accepted after the work completed in this research, the CNN models created are not perfect at performing their tasks. Therefore a number of examples of images that were misclassified from an age and

Age Classification	Exact	One-off
Levi & Hassncer	50.7%	84.7%
This research	72.2%	90.8%

Table 4.9: Exact versus one-off accuracy - compared against Levi & Hassncer, 2015

gender point of view have been provided below. Looking at potential reasons for the misclassification, in Figure 4.5(a) below where the male was classified as female, this image contains clothing that is covering essential features of gender identity that could potentially be the reason for the misclassification. The subject is wearing a hat and hood covering the subjects hair completely while the jacket the subject is wearing is also partially covering the subjects lower face. It is possible that the excessive clothing in this scenario is the reason for the misclassification. In the second example, Figure 4.5(b) the subject was classified as male when they are in fact female. In this example, it's possible the misclassification is due to the age of the subject in the image. The subject falls within the age bracket of 4-6 and it is possible that the youthful appearance of the subject has been the cause of the incorrect classification being applied.

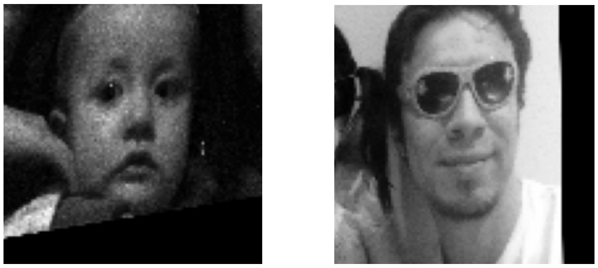


(a) Male classified as female

(b) Female classified as male

Figure 4.5: Examples of incorrectly classified genders

Moving on to look at examples where the incorrect age of a subject was applied. Looking at the first figure, Figure 4.6(a) in which the 0-2 year old was classified as someone who is between the age of 48-53. And while this is certainly a large misclassification, after looking at the image it is possible to see why the CNN misclassified by such a large extent. The image itself is a rather dull image with plenty of shadows present. There is also a lot of noise in terms of image blurriness and the image itself cuts off at an angle in the bottom right hand corner. When each of these issues are combined, it is possible that it made the image very difficult to decipher in terms of age for the CNN. Next looking at Figure 4.6(b) in which the subject is between the age of 25-32 but was classified in the age bracket 15-20. This difference is less significant in comparison to figure 4.6(a) with the 15-20 age bracket being the age bracket directly below the 25-32 age bracket. And looking at the image, the stand out feature that could possibly have led to the misclassification is the sunglasses being worn by the subject. The sunglasses are rather large and cover the subjects eyes completely making it impossible to see any potential ageing in the eye area that typically comes with ageing from 15-20 to 25-32. A point of interest would be to see what age bracket this subject would be placed in if an image without sunglasses existed.



(a) 0-2 year old classified as 48-53

(b) 25-32 year old classified as 15-20

Figure 4.6: Examples of incorrectly classified genders

The results of this research have shown that the use of both standard and novel augmentation techniques can have a largely beneficial impact on the area of image classification. However that is not to say that further improvements could not be made through future work or that there are no issues with this research currently. A stand out issue that could be fixed with the right resources is the environment in which these models were created in. Google Colab is a fantastic free service that allows users to create excellent, highly performing models. However limitations such as RAM and session timeout issues are unfortunate issues that cannot be removed.

Colab comes with 13GB of RAM as standard which is more than enough for standard tasks. However, in this research it was originally intended to augment the dataset even further in an attempt to improve on the results even more. However this was not possible as any more images above the current augmented number, approximately 70,000, resulted in the Colab session crashing during CNN training due to lack of RAM. Also as previously mentioned, while attempting to train an SVM with the augmented dataset, the Colab cell running time limit was breached resulting in not being able to create an SVM trained with the fully augmented dataset. Both these problems could be solved with better resources that contain more RAM and do not contain any cell running time limit.

Another issue that cannot be ignored but could be removed in future work revolves around the use of Faceapp to augment the data. One of the key benefits of Faceapp is that it offers both a free and premium version of the application. The main difference between the two is that there are more filters available in the premium version that allow the user to apply many different filters other than ageing filters. However it is not the lack of filters that is the issue in this circumstance but the inclusion of a watermark that Faceapp applies to altered photos when using the free version of the application. As can be seen in the example photo in the Results section, the word ”Faceapp” is placed in the bottom right corner of the image. This is the case for all images in the dataset that were augmented with the Faceapp application. This results in images with the same pattern of pixel brightness in the bottom right corner. The issue with this is that when the model was training it may have potentially learned to familiarise itself this brightness in the bottom right and associated it with images of subjects in the 60-100 age bracket. A counter argument to this is that there were more images in the 60-100 age bracket that had not been processed through Faceapp than those that had and therefore it is possible that the model learned to identify subjects within this age bracket sufficiently, regardless of the watermark. A potential solution that could be employed for any future work in this area using Faceapp would be to use the premium version the application as this version does not place the watermark on processed images.

A slightly less impacting issue that comes into play with the Faceapp application revolves around its gender classification capabilities. When an image is uploaded to Faceapp, the application automatically makes an attempt at classifying the gender of the person in the image. This is done, as depending on the gender, different filter options are provided for the user to apply to the image. An important feature of the application to point out is that it applies the ageing filter to images slightly different depending on whether the image of the person is male or female. For example with males it tends to recede the hairline of the subject slightly while with females this is not the case. And while Faceapp is an excellent application capable of identifying the correct gender of an image subject with a high degree of accuracy, it did not always correctly classify the gender of the subjects in the Adience dataset. Therefore in cases where Faceapp incorrectly classified the gender of the image subject, the ageing filter was not applied. In the grand scheme of this research, this problem is not a major one as it is just a case of not including the image in the Faceapp augmentation process, however it is worth pointing out in case an attempt at replicating this research was undertaken or if this research was to be re-run.

Additionally to the issues mentioned above, another one exists with the use of the image rotation augmentation technique. As can be seen in the example images in the results section, by rotating the images they are essentially being cropped at each corner resulting in areas of total black pixels. In an ideal scenario, the image would still fit the 100 x 100 shape so as not to leave these black pixel areas. These pixel areas could have potentially effected the model results in a negative manner as it would have been more difficult for the model to determine the age and gender of the subjects in these images as they would all have had this same consistent black pixel area. One of the ways this issue could have been resolved would have been to zoom in on the image before rotating it but this would bring it's own challenges. By zooming in on the image first it would have insured that when it was then rotated, that none of the corners would have been cut off. However, had zooming been applied to the images it would have resulted in images where the entire photo was no longer visible and potentially cases where not all of the subject in the image was visible. Therefore the model would have been attempting to learn the ages and genders of subjects that are not fully visible in the image which could have potentially have had negative impacts on the model results also.

A final issue with this research, and any previous research that has used this Adience dataset, again falls within the area of age classification. Because of the difficulty of classifying a human to an exact year of age, this dataset attempts to overcome this by binning the years into age brackets. And while this is an effective means of making the classification problem easier to complete, the age brackets implemented do not cover all years of age. For example one of the age brackets is 15-20 years old while the next age bracket is 25-32. In this scenario the ages of 21-24 are left unaccounted for with no representation in the dataset. This issue has already been addressed by the creators of this dataset, Eidinger et al. (2014), with the reason for certain years of age being excluded as it is very difficult, even for humans, to determine the exact age of a person. And in the context of this dataset, a person who is 20 years old would still look almost identical at 21 years old. Therefore if all ages were to be accounted for, it would make the already difficult task of age classification even more difficult as a CNN would find it extremely tricky to classify whether a person was 20 or 21.

Regardless of the issues mentioned above this research has shown that data augmentation can have beneficial and often a significant impact on the performance of CNNs in an image classification task. With further improvements in AI photo editing technology such as that of the Faceapp application it may be possible to see even further significant improvements. This is particularly true in the area of age classification which still has a long way to go before becoming a seamless and full-proof process. In an ideal world, Faceapp or a similar application would allow for the batch uploading of images to make the process of applying an age filter a much less manual job. If this point is ever reached, this research has shown that an application like Faceapp can improve age and gender classification models on a much larger scale than that of this research.

Chapter 5

Conclusion

5.1 Research Overview

The area of age and gender classification has become a widely popular area of study in the past 20 years. With the increasing use of biometric information in our everyday lives, a price cannot be put on the benefit of improving these type of classification models. Age and gender classification is a task that can be completed in many different forms. Convolutional neural networks have been a very popular method in recent times however research completed by Sedaaghi, M. H. (2009) and Kaushik et al. (2019) show that different model and data types can be used and can often lead to better results than CNNs can currently reach. Previous research has shown excellent results through the use of different models such as CNNs, SVMs, DNNs, and HyperBF networks to name a few while also using data other than facial images such as speech and audio data or brain wave activity data. And while this age and gender classification topic is a well researched area, in the CNN space there appeared to be an opportunity to expand on previous research through the use of novel data augmentation techniques, particularly in the age classification side of things.

5.2 Problem Definition

From researching many different papers in this area, a common theme was noticed that a lack of data was often noted as the reason for results not reaching the desired level or that the introduction of more augmentation techniques would be included in future work. In particular, it was noted that an AI image editor had never been used as a means of creating more data from which the CNN could learn from. It was because of this finding that the use of an AI image editor, in the shape of Faceapp, was used as a means of data augmentation. Faceapp was the application of choice due to it's excellent reviews from the general public on the Apple App and Google Play Store along with the ease of exporting the filter processed images from the application. The purpose of Faceapp was to help reduce the problem faced in many age and gender classification tasks which was the difficulty seen in classifying a human to a particular age group. Gender classification had seen relatively high levels of success in CNN focused research papers. In this research the increasing of the dataset size through image mirroring and image rotation further improved on the gender classification results. On the other hand in other research papers, age classification results were typically far lower than gender classification results due in part to a lack of data to train the models on. To combat this issue, Faceapp was used to augment the dataset by applying a filter to the subjects of images to make them appear older than they actually are - thus creating additional versions of the images. In particular it was used to increase the dataset size of subjects in the over 60 age bracket. This age bracket was represented far less in the Adience dataset than younger age groups. Therefore Faceapp was employed to increase the population in an attempt to further balance the dataset. And with the results achieved in this research, it appears the use of Faceapp to augment the dataset provided an improvement to the model and could potentially be used on a larger scale in future research papers in this age classification space.

5.3 Design/Experimentation, Evaluation & Results

A number of CNN and SVM models were created as part of this research in order to determine the best model type for performing age and gender classification on the Adience dataset. After creating both model types it was clear that the CNN was the more successful model achieving an accuracy of 86.9% for gender classification and 65.1% accuracy for age classification. These two CNN models significantly outperformed their SVM counterparts for this initial task. The next and more important challenge was to analyse the impact that data augmentation could have on the CNN models accuracy results. Having researched the topic of data augmentation and gaining an understanding that it typically helps to improve CNN results, the idea was to try and decide upon a novel approach to integrate into the study. Standard augmentation approaches such as image mirroring and image rotating have been used in multiple studies in the past and are often the go to techniques for researchers looking for more data. Because these techniques had improved results in past research papers they were an obvious choice to include first. It was the introduction of Faceapp as a means of data augmentation that was a new approach that had not been seen or used in the past. The idea was to see if Faceapp augmentation could aid in improving the model accuracy results. Fortunately, this appeared to be the case with an improvement seen in the model from 65.1% to 72.2%. However, because image mirroring and image rotating techniques had also been applied to the dataset, a decision was made to compare the results of the CNN trained on the dataset that had image mirroring and rotating applied versus a CNN trained on the dataset that had both of those techniques plus the Faceapp augmentation applied also. This decision was made to ensure that it wasn't just the image mirroring and rotating that were the cause of the improvement. Interestingly, the model trained using the dataset that only had the image mirroring and rotating applied performed better than the CNN model with no augmentation but worse than the model that had image mirroring, image rotating, and Faceapp augmentation applied. This model achieved an accuracy of 69.6% in comparison to the best performing models accuracy of 72.2%. The improvement Faceapp made to the model may not be substantial, but in an area such as age classification that is already so difficult to do, this improvement could be very significant for future research.

5.4 Contributions and impact

The importance of improving models for age, gender, and general human image classification cannot be understated. These model types have a vast amount of uses in the real world today and improving on the results achieved by these models can only expand their use cases. This research has shown a novel means of augmenting datasets through the use of Faceapp and has shown promise as an effective means of doing so with the improvement of the CNN results. In the future, with continued improvement, models like these could be applied to a number of different sectors, bringing different benefits to each. An example of sectors could be the law enforcement sector. A model such as this could help with the classification of a suspect or victim, aiding in any ongoing investigations. Other scenarios of use for CNN models such as those created in this research are the healthcare sector. These models could be implemented in a health care scan scenario as an additional aid in identifying illnesses or prescription of treatment. Additionally, these CNN models could have use in the security sector. The use of biometrics in the security industry is ever on the rise with biometric information being used to unlock our personal devices and also in an international travelling capacity such as in airports. Models such as those created in this research study could be implemented in similar setting such as in corporate offices for gaining access to buildings or in the pub and nightclub industry as a form of identification process. A final, and very interesting area these models could be applied is the area of targeted advertising through social media. A potential solution could be used where images are analysed from users social media profiles, and depending on the age and gender of the classified user, specific advertisements could be targeted towards that user. This solution could be a potential game-changer and a very simple means of targeting wide demographic groups using the freely provided image data. In order for these models to be implemented in real world scenarios such as those mentioned above, the accuracy of them will need to improve. However with novel augmentation techniques continuously helping improve the accuracy of CNNs, it may not be long before we see them in use on a daily basis.

5.5 Future Work & recommendations

In order for an application like Faceapp to be used in future work in a large scale manner for age and gender classification the ease of applying filters to images would need to be improved. Currently, the process of applying filters to images is a very manual process with filters having to be applied to images one after another by the user of the application. If a bulk upload system could be created by the Faceapp developers, or any other developers of similar applications, the usefulness of the application could increase ten fold. Faceapp as an application offers a number of other filter types such as the ability to swap gender or to artificially make the subject change facial expression such as making the subject of the image appear to be smiling. From testing, these filters do not appear as realistic to the human eye as the ageing filter. However, with improvements, these filters could offer more novel means of data augmentation which we have seen can be beneficial in the age and gender classification space. For future work, if Faceapp could implement this image bulk upload system and improve on other filter types this type of research could be taken on to another level. Additional ideas for future work would be the use of these CNN models on group photos rather than images of individuals. In an ideal world, the CNNs would be capable of recognising multiple faces in an image and therefore making multiple classifications depending on the number of people in the photo.

References

Agrawal, B., & Dixit, M. (2020). Age Estimation and Gender Prediction Using Convolutional Neural Network (pp. 163–175). Springer, Cham. https://doi.org/10.1007/978-3-030-44758-8_15

Al-Azzawi, D. S. (2019). Human Age and Gender Prediction Using Deep Multi-Task Convolutional Neural Network. Journal of Southwest Jiaotong University, 54(4). https://doi.org/10.35741/issn.0258-2724.54.4.11

Antipov, G., Baccouche, M., Berrani, S. A., & Dugelay, J. L. (2017). Effective training of convolutional neural networks for face-based gender and age prediction. Pattern Recognition, 72, 15–26. https://doi.org/10.1016/j.patcog.2017.06.031

Antipov, G., Berrani, S. A., & Dugelay, J. L. (2016). Minimalistic CNN-based ensemble model for gender prediction from face images. Pattern Recognition Letters, 70, 59–65. https://doi.org/10.1016/j.patrec.2015.11.011

Bedeli, M., Geradts, Z., & van Eijk, E. (2018). Clothing identification via deep learning: forensic applications. Forensic Sciences Research, 3(3), 219–229. https://doi.org/10.1080/20961790.2018.1526251

Binary Cross Entropy/Log Loss for Binary Classification. (2021). Retrieved May 27, 2021, from https://www.analyticsvidhya.com/blog/2021/03/binary-cross-entropy-log-loss-for-binary-classification/

Bipembi, H., ben Hayfron-Acquah, J., Kobina Panford, J., Appiah, O., Hayfron-Acquah, J. B., & Panford, J. K. (2015). Calculation of Body Mass Index using Image Processing Techniques PhD thesis View project Parity Progression estimation and categorical Analysis of Birth cohort data in ghana View project Calculation of Body Mass Index using Image Processing Techniques. In International Journal of Artificial Intelligence and Mechatronics (Vol. 4, Issue 1). https://www.researchgate.net/publication/280133090

Dehghan, A., Ortiz, E. G., Shu, G., & Masood, S. Z. (2017). DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Network. ArXiv. http://arxiv.org/abs/1702.04280

Donati, L., Iotti, E., Mordonini, G., & Prati, A. (2019). Fashion Product Classification through Deep Learning and Computer Vision. Applied Sciences, 9(7), 1385. https://doi.org/10.3390/app9071385

Fazl-Ersi, E., Mousa-Pasandi, M. E., Laganiere, R., & Awad, M. (2014). Age and gender recognition using informative features of various types. 2014 IEEE International Conference on Image Processing, ICIP 2014, 5891–5895. https://doi.org/10.1109/ICIP.2014.7026190

Günay, A., & NabIyev, V. v. (2008). Automatic age classification with LBP. 2008 23rd International Symposium on Computer and Information Sciences, ISCIS 2008. https://doi.org/10.1109/ISCIS.2008.4717926

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Han, D., Liu, Q., & Fan, W. (2018). A new image classification method using CNN transfer learning and web data augmentation. Expert Systems with Applications, 95, 43–56. https://doi.org/10.1016/j.eswa.2017.11.028

Horng, W.-B., Lee, C.-P., & Chen, C.-W. (2001). Classification of Age Groups Based on Facial Features. In Tamkang Journal of Science and Engineering (Vol. 4, Issue 3).

Hussain, Z., Gimenez, F., Yi, D., & Rubin, D. (2017). Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. AMIA ... Annual Symposium Proceedings. AMIA Symposium, 2017, 979–984. /pmc/articles/PMC5977656/

Ito, K., Kawai, H., Okano, T., & Aoki, T. (2019). Age and Gender Prediction from Face Images Using Convolutional Neural Network. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings, 7–11. https://doi.org/10.23919/APSIPA.2018.8659655

K, R., K, R., Raja, K. B., R, V. K., & Patnaik, L. M. (n.d.). Feature Extraction based Face Recognition, Gender and Age Classification. Retrieved December 18, 2020, from http://citeseerx.ist.psu.edu/viewdoc/

Kaushik, P., Gupta, A., Roy, P. P., & Dogra, D. P. (2019). EEG-Based Age and Gender Prediction Using Deep BLSTM-LSTM Network Model. IEEE Sensors Journal, 19(7), 2634–2641. https://doi.org/10.1109/JSEN.2018.2885582

Kumar, V. (2020). Age Prediction using Image Dataset using Machine Learning. International Journal of Innovative Technology and Exploring Engineering, 8(12s3), 107–113. https://doi.org/10.35940/ijitee.L1020.10812S319

Kwon, Y. H., & da Vitoria Lobo, N. (1999). Age classification from facial images. Computer Vision and Image Understanding, 74(1), 1–21. https://doi.org/10.1006/cviu.1997.0549

Levi, G., & Hassncer, T. (2015). Age and gender classification using convolutional neural networks. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2015-October, 34–42. https://doi.org/10.1109/CVPRW.2015.7301352

Li, W., Wu, G., Zhang, F., & Du, Q. (2017). Hyperspectral Image Classification Using Deep Pixel-Pair Features. IEEE Transactions on Geoscience and Remote Sensing, 55(2), 844–853. https://doi.org/10.1109/TGRS.2016.2616355

Moghaddam, B., & Yang, M. H. (2000). Gender classification with support vector machines. Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000, 306–311. https://doi.org/10.1109/AFGR.2000.840651

Nagi, J., Khaleel, S., & Nagi, A. F. (n.d.). A MATLAB based Face Recognition System using Image Processing and Neural Networks.

Ng, C. B., Tay, Y. H., & Goi, B. M. (2014). Comparing image representations for training a convolutional neural network to classify gender. Proceedings - 1st International Conference on Artificial Intelligence, Modelling and Simulation, AIMS 2013, 29–33. https://doi.org/10.1109/AIMS.2013.13

Poggio, B., Brunelli, R., & Poggio, T. (1995). HyberBF Networks for Gender Classification. Retrieved December 18, 2020, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.2814

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Age and gender classification from speech and face images by jointly fine-tuned deep neural networks. Expert Systems with Applications, 85, 76–86. https://doi.org/10.1016/j.eswa.2017.05.037

Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). An All-In-One Convolutional Neural Network for Face Analysis. Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, 17–24. https://doi.org/10.1109/FG.2017.137

Rattani, A., Reddy, N., & Derakhshani, R. (2018). Convolutional neural networks for gender prediction from smartphone-based ocular images. IET Biometrics, 7(5), 423–430. https://doi.org/10.1049/iet-bmt.2017.0171

Rizvi, M. S. Z. (2020, February 18). CNN Image Classification — Image Classification Using CNN. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/

Safak, E., & Bariicc, N. (2018, December 6). Age and Gender Prediction Using Convolutional Neural Networks. ISMSIT 2018 - 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings. https://doi.org/10.1109/ISMSIT.2018.8567066

Sedaaghi, M. H. (2009). A Comparative Study of Gender and Age Classification in Speech Signals. In Iranian Journal of Electrical & Electronic Engineering (Vol. 5, Issue 1). IRANIAN JOURNAL OF ELECTRICAL AND ELECTRONIC ENGINEERING. www.SID.ir

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). Research on data augmentation for image classification based on convolution neural networks. Proceedings - 2017 Chinese Automation Congress, CAC 2017, 2017-January, 4165–4170. https://doi.org/10.1109/CAC.2017.8243510

Srinivas, N., Atwal, H., Rose, D. C., Mahalingam, G., Ricanek, K., & Bolme, D. S. (2017). Age, Gender, and Fine-Grained Ethnicity Prediction Using Convolutional Neural Networks for the East Asian Face Dataset. Proceedings - 12th IEEE International Conference on Automatic Face and Gesture Recognition, 953–960. https://doi.org/10.1109/FG.2017.118

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1891–1898. https://doi.org/10.1109/CVPR.2014.244

'Uyun, S., & Efendi, T. (2019). Classification of Human Weight Based on Image. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), 13(2), 105. https://doi.org/10.22146/ijccs.35794

Wang, C., & Xi, Y. (n.d.). Convolutional Neural Network for Image Classification.

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, 6382–6388. http://arxiv.org/abs/1901.11196

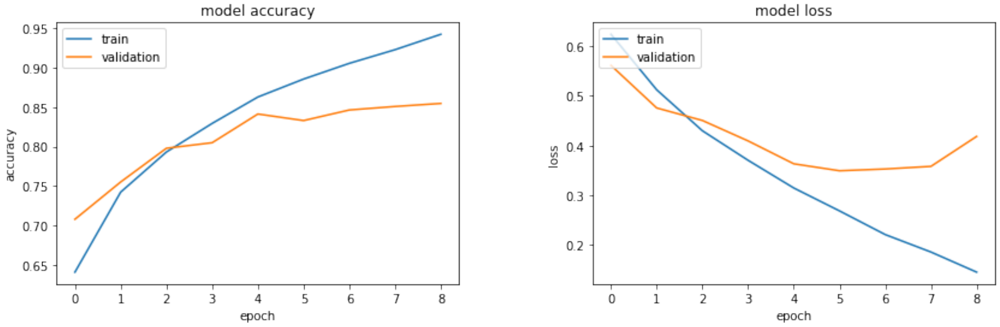
Wigington, C., Stewart, S., Davis, B., Barrett, B., Price, B., & Cohen, S. (2017). Data Augmentation for Recognition of Handwritten Words and Lines Using a CNN-LSTM Network. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR, 1, 639–645. https://doi.org/10.1109/ICDAR.2017.110

Yamaguchi, O., Fukui, K., & Maeda, K. I. (1998). Face recognition using temporal image sequence. Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998, 318–323. https://doi.org/10.1109/AFGR.1998.670968

Yu, S., Tan, T., Huang, K., Jia, K., & Wu, X. (2009). A study on gait-based gender classification. IEEE Transactions on Image Processing, 18(8), 1905–1910. https://doi.org/10.1109/TIP.2009.2020535

Appendix A

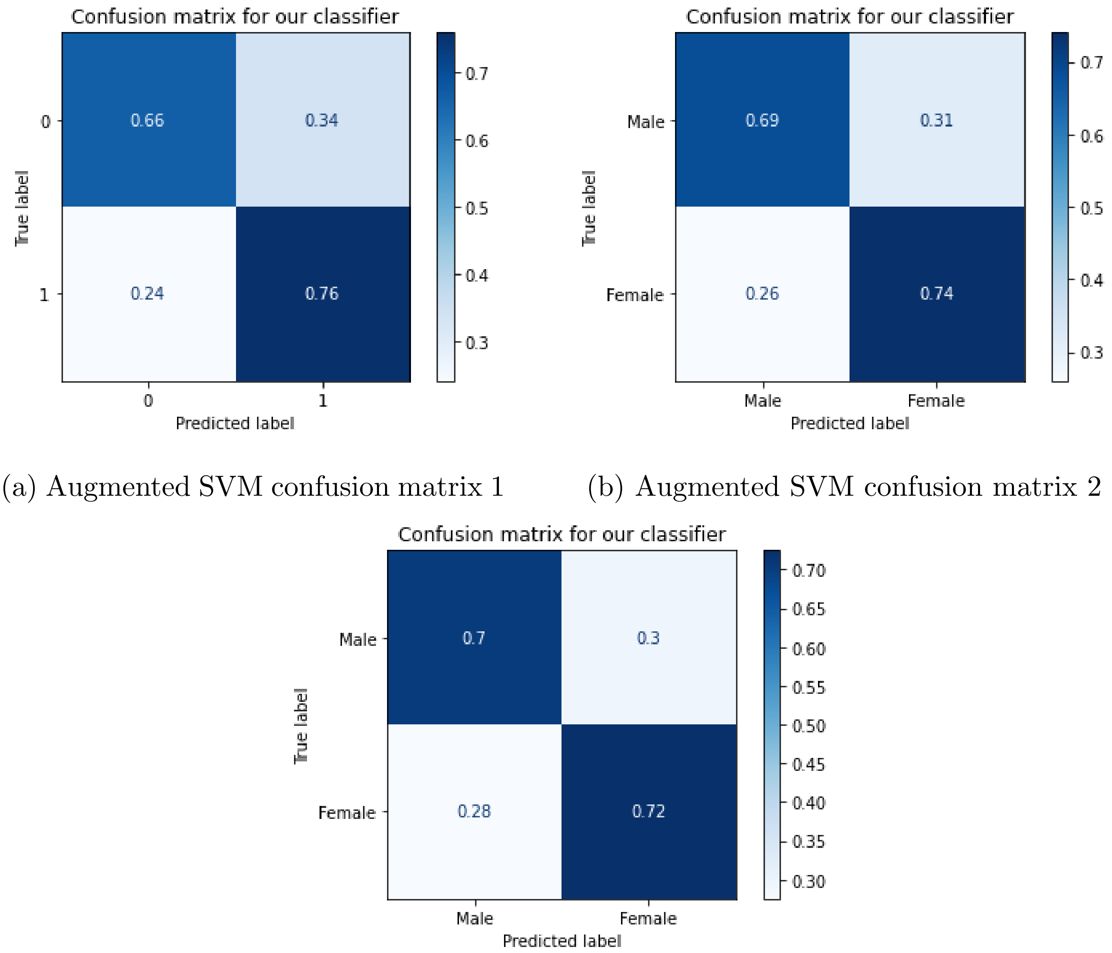
Additional content



(a) Gender CNN Accuracy

(b) Gender CNN Loss

Figure A.1: Examples of accuracy and loss charts for Gender Classifying CNN

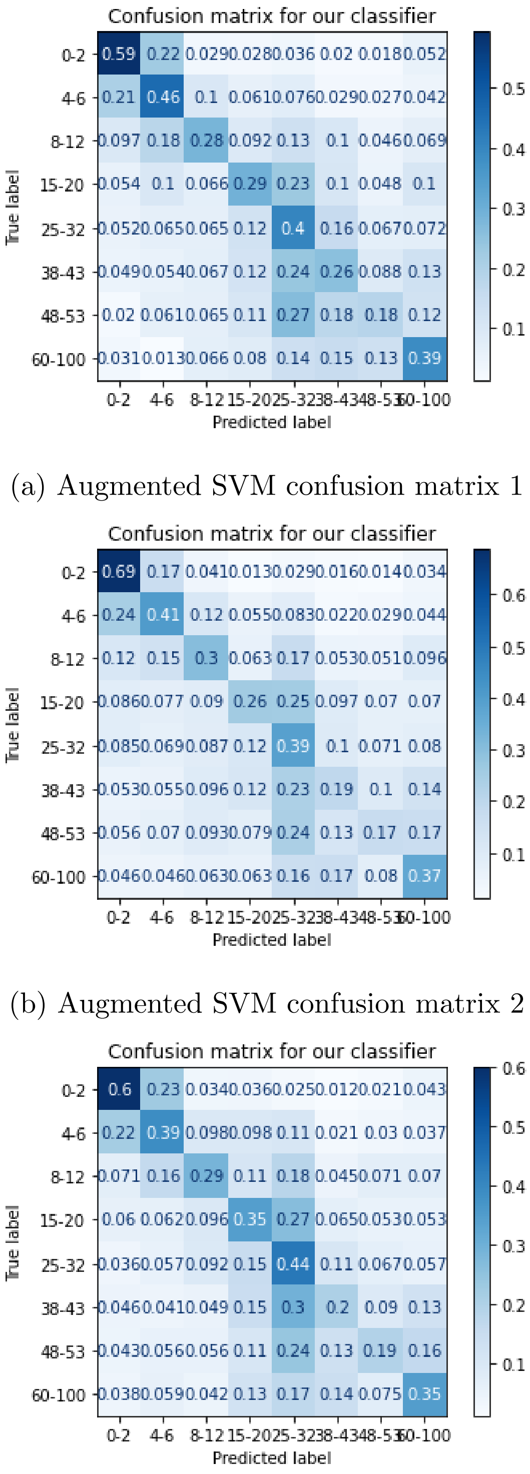


(a) Augmented SVM confusion matrix 1

(b) Augmented SVM confusion matrix 2

(c) Augmented SVM confusion matrix 3

Figure A.2: Confusion matrices for the 3 SVM models trained on the split augmented dataset - Gender



	0-2	4-6	8-12	15-20	25-32	38-43	48-53	60-100
0-2	0.59	0.22	0.029	0.028	0.036	0.02	0.018	0.052
4-6	0.21	0.46	0.1	0.061	0.076	0.029	0.027	0.042
8-12	0.097	0.18	0.28	0.092	0.13	0.1	0.046	0.069
15-20	0.054	0.1	0.066	0.29	0.23	0.1	0.048	0.1
25-32	0.052	0.065	0.065	0.12	0.4	0.16	0.067	0.072
38-43	0.049	0.054	0.067	0.12	0.24	0.26	0.088	0.13
48-53	0.02	0.061	0.065	0.11	0.27	0.18	0.18	0.12
60-100	0.031	0.013	0.066	0.08	0.14	0.15	0.13	0.39

(a) Augmented SVM confusion matrix 1

	0-2	4-6	8-12	15-20	25-32	38-43	48-53	60-100
0-2	0.69	0.17	0.041	0.013	0.029	0.016	0.014	0.034
4-6	0.24	0.41	0.12	0.055	0.083	0.022	0.029	0.044
8-12	0.12	0.15	0.3	0.063	0.17	0.053	0.051	0.096
15-20	0.086	0.077	0.09	0.26	0.25	0.097	0.07	0.07
25-32	0.085	0.069	0.087	0.12	0.39	0.1	0.071	0.08
38-43	0.053	0.055	0.096	0.12	0.23	0.19	0.1	0.14
48-53	0.056	0.07	0.093	0.079	0.24	0.13	0.17	0.17
60-100	0.046	0.046	0.063	0.063	0.16	0.17	0.08	0.37

(b) Augmented SVM confusion matrix 2

	0-2	4-6	8-12	15-20	25-32	38-43	48-53	60-100
0-2	0.6	0.23	0.034	0.036	0.025	0.012	0.021	0.043
4-6	0.22	0.39	0.098	0.098	0.11	0.021	0.03	0.037
8-12	0.071	0.16	0.29	0.11	0.18	0.045	0.071	0.07
15-20	0.06	0.062	0.096	0.35	0.27	0.065	0.053	0.053
25-32	0.036	0.057	0.092	0.15	0.44	0.11	0.067	0.057
38-43	0.046	0.041	0.049	0.15	0.3	0.2	0.09	0.13
48-53	0.043	0.056	0.056	0.11	0.24	0.13	0.19	0.16
60-100	0.038	0.059	0.042	0.13	0.17	0.14	0.075	0.35

(c) Augmented SVM confusion matrix 3

Figure A.3: Confusion matrices for the 3 SVM models trained on the split augmented dataset - Age