PhD Thesis

# Digital Forensics: Leveraging Deep Learning Techniques in Facial Images to Assist Cybercrime Investigations

Felix Santiago Anda Basabe, MSc, BEng

A report submitted in part fulfilment of the degree of

**PhD in Computer Science**

**Supervisors:**

Dr Mark Scanlon & Dr Nhien-An Le-Khac

**Head of School:** Assoc. Prof. Chris Bleakley

UCD School of Computer Science
College of Science
University College Dublin

13th May 2021

# CONTENTS

**List of Figures** xi

**List of Tables** xiii

**List of Algorithms** xiv

**List of Code Snippets** xv

**List of Abbreviations** xvi

**Abstract** xix

**Statement of Original Authorship** xx

**Collaborations** xxi

**Acknowledgements** xxii

**Dedication** xxiii

**List of Publications** xxv

**1 Introduction** **1**
1.1 Preamble . . . . . 1
1.2 Research Ethics . . . . . 3
1.3 Research Questions . . . . . 4
1.4 Contributions . . . . . 4
1.5 Layout of this Thesis . . . . . 5

**2 Background Research** **7**
2.1 Cybercrime and Digital Forensic Investigations . . . . . 7
2.1.1 Cybercrime . . . . . 7
2.1.1.1 Introduction . . . . . 7
2.1.1.2 Challenges . . . . . 8
2.1.1.3 Classification of Cybercrime . . . . . 9
2.1.2 Digital Forensics . . . . . 12

2.1.2.1 Introduction . . . . . 12
2.1.2.2 Digital Forensic Process . . . . . 12
2.1.2.3 Process Models Applicable to CSEM . . . . . 14
2.1.2.4 Admissibility of CSEM Evidence . . . . . 16
2.1.2.5 Existing Tools and Techniques . . . . . 17
2.1.2.6 Digital Forensic Challenges . . . . . 18
2.1.2.7 Digital Forensic Backlog . . . . . 19
2.1.2.8 Triage . . . . . 20
2.2 Child Sexual Exploitation Material Investigation . . . . . 21
2.2.1 Introduction . . . . . 21
2.2.2 Prevalence . . . . . 22
2.2.3 Cost . . . . . 23
2.2.4 Terminology Used in the Literature . . . . . 23
2.2.5 Modus-Operandi . . . . . 24
2.2.6 Approaches to Identify and Combat CSEM . . . . . 26
2.2.6.1 Traditional Hash Database Approach . . . . . 26
2.2.6.2 Improved Hash Search Approach . . . . . 27
2.2.6.3 Machine Learning Approach . . . . . 28
2.2.7 Identification of Subjects via Visual Media . . . . . 29
2.3 Dataset Curation . . . . . 30
2.3.1 Data Selection . . . . . 31
2.3.2 Data Collection . . . . . 31
2.3.3 Data Organisation . . . . . 31
2.3.4 Dataset Evaluation and Validation . . . . . 32
2.3.5 Data Storage . . . . . 33
2.4 Machine Learning . . . . . 33
2.4.1 Introduction . . . . . 33
2.4.2 Machine Learning Tasks . . . . . 35
2.4.2.1 Classification . . . . . 35
2.4.2.2 Regression . . . . . 35
2.4.2.3 Anomaly Detection . . . . . 35
2.4.3 Generalisation . . . . . 36
2.4.4 Evaluation Model Metrics . . . . . 37
2.4.5 Optimisation Strategies . . . . . 38
2.4.5.1 Curriculum Learning . . . . . 38
2.4.5.2 Knowledge Distillation . . . . . 38
2.4.5.3 Supervised Pre-training and Transfer Learning . . . . . 39
2.4.6 Regularisation for Deep Learning . . . . . 40
2.4.6.1 Dataset Augmentation Techniques . . . . . 40
2.4.6.2 Early Stopping . . . . . 41
2.4.6.3 Ensemble Learning . . . . . 41
2.4.6.4 Generative Adversarial Networks . . . . . 41
2.4.7 Evaluation Protocol . . . . . 42
2.4.7.1 Dataset Pre-Processing . . . . . 42
2.4.7.2 Dataset Split . . . . . 42
2.4.8 Network Architecture . . . . . 43
2.4.9 Hyper-Parameters . . . . . 45

2.4.9.1 Learning Rate . . . . . 45
2.4.9.2 Optimisation . . . . . 45
2.4.10 Artificial Intelligence for Digital Forensics . . . . . 46
2.4.11 Object and Face Recognition . . . . . 47

**3 Related Work** **48**
3.1 Face Recognition . . . . . 48
3.1.1 Introduction . . . . . 48
3.1.2 Facial Features . . . . . 48
3.1.3 Face Detection . . . . . 50
3.1.3.1 Offline . . . . . 50
3.1.3.2 Online . . . . . 51
3.2 Soft Biometrics . . . . . 52
3.2.1 Introduction . . . . . 52
3.2.2 Age Estimation . . . . . 53
3.2.3 Human Facial Age Perception . . . . . 53
3.2.4 Facial Age Estimation . . . . . 54
3.2.4.1 Offline Models . . . . . 54
3.2.4.2 Online Models . . . . . 56
3.2.5 Facial Vector Embeddings and Soft Biometric Traits . . . . . 56
3.2.6 Regression vs Classification . . . . . 58
3.2.7 Underage Facial Age Estimation . . . . . 58
3.2.8 Gender Estimation . . . . . 59
3.2.9 Skin Detection and Skin Tone Classifiers . . . . . 59
3.2.10 Race/Ethnicity Prediction . . . . . 60
3.2.11 Influencing Factors on Age Estimation . . . . . 61
3.2.11.1 Facial Expressions . . . . . 61
3.2.11.2 Noise . . . . . 61
3.2.11.3 Makeup . . . . . 61
3.2.11.4 Gender . . . . . 61
3.2.11.5 Ethnicity . . . . . 62
3.3 Datasets for Facial Age Estimation . . . . . 63
3.3.1 Introduction . . . . . 63
3.3.2 Small Datasets . . . . . 63
3.3.2.1 FG-NET . . . . . 63
3.3.2.2 MEDS I & II . . . . . 65
3.3.2.3 FERET . . . . . 66
3.3.3 Medium Datasets . . . . . 66
3.3.3.1 Adience . . . . . 66
3.3.3.2 UTKFace . . . . . 67
3.3.4 Large Datsets . . . . . 69
3.3.4.1 Selfie-FV . . . . . 69
3.3.4.2 MORPH . . . . . 70
3.3.4.3 IMDB-WIKI . . . . . 72
3.3.4.4 Yahoo Flickr Creative Commons 100M . . . . . 73
3.3.5 Underage Facial Age Datasets . . . . . 73
3.3.6 Data Pre-Processing Stage . . . . . 74

3.4 Data Bias and Ethical Controversy . . . 75
3.5 Summary of Literature Review . . . 76

**4 Methodology** **78**
4.1 Establishing a Baseline for Facial Age Estimation Accuracy . . . 79
4.1.1 Offline Models as a Baseline . . . 80
4.1.2 Online Services as Resources . . . 80
4.1.3 Grant and Resource Access Timeline for Cloud-Based Instances 82
4.1.4 Datasets with Age Labels for Bench-marking . . . 83
4.2 Dataset Generation - Standalone . . . 85
4.2.1 Overview of the Software Solution for Performance Evaluation 85
4.2.2 Design Patterns . . . 85
4.2.3 Image Database Design . . . 87
4.2.4 Proposed Architecture . . . 87
4.2.5 Evaluation Performance Methodology . . . 89
4.2.6 Service Database Design . . . 89
4.3 Dataset Generation - Online . . . 90
4.3.1 Ethical Considerations . . . 91
4.3.1.1 Access to Data . . . 91
4.3.1.2 Creative Commons and GDPR . . . 91
4.3.1.3 Anonymity & Disclosure . . . 92
4.3.2 VisAGe Design/Methodology . . . 93
4.3.2.1 Photo Acquisition . . . 94
4.3.2.2 Categorisation . . . 95
4.3.2.3 Image Selection Criteria . . . 95
4.3.2.4 Voter Background . . . 95
4.3.2.5 Voting Procedure . . . 96
4.3.2.6 Database Design . . . 96
4.4 Dlib Contour Artistic Approach for Facial Image Pre-processing . . . 97
4.4.1 Introduction . . . 97
4.4.2 Facial Image Pre-processing . . . 98
4.4.3 DCA Facial Proportion Artistic Approach . . . 99
4.5 Underage Age Estimation Model Design . . . 102
4.5.1 DS13K . . . 102
4.5.1.1 Dataset Curation . . . 103
4.5.1.2 Dataset Pre-Processing . . . 103
4.5.1.3 Dataset Split . . . 104
4.5.1.4 Dataset Augmentation . . . 104
4.5.1.5 Network Architecture . . . 104
4.5.1.6 Hyper-Parameters . . . 105
4.5.2 DeepUAge . . . 105
4.5.2.1 Dataset Curation . . . 105
4.5.2.2 Dataset Pre-Processing . . . 107
4.5.2.3 Dataset Split . . . 107
4.5.2.4 Dataset Augmentation . . . 107
4.5.2.5 Network Architecture . . . 108
4.5.2.6 Hyper-parameters . . . 110

4.5.3 Vec2UAge . . . 110

4.5.3.1 Dataset Curation . . . 110

4.5.3.2 Dataset Pre-Processing . . . 112

4.5.3.3 Dataset Split . . . 112

4.5.3.4 Data Augmentation . . . 112

4.5.3.5 Network Architecture . . . 113

4.5.3.6 Hyper-Parameters . . . 114

4.6 Influencing Factors on the Accuracy of Automated Underage Facial Age Estimation . . . 115

4.6.1 Cloud Services . . . 115

4.6.1.1 Microsoft Azure: Face API . . . 116

4.6.1.2 Amazon AWS: Rekognition Service . . . 116

4.6.2 Pearson Correlation Coefficient . . . 116

**5 Implementation** **118**

5.1 Dataset Generator . . . 118

5.1.1 Database Superclass . . . 118

5.1.2 Dataset Plugins . . . 119

5.1.3 NoSQL Dataset Implementation . . . 120

5.1.4 Dataset Generator GUI . . . 121

5.2 Evaluation Platform . . . 122

5.2.1 Service Superclass . . . 122

5.2.2 Service Plugins . . . 124

5.2.3 NoSQL Service Implementation . . . 124

5.2.4 Service Generator GUI . . . 125

5.2.5 Plots . . . 125

5.3 VisAGe - Dataset Framework . . . 126

5.3.1 Technical Overview . . . 127

5.3.1.1 Background Scripts . . . 127

5.3.1.2 Data Records . . . 130

5.3.1.3 Web Applications . . . 132

5.4 Dlib Contour Artistic Approach . . . 133

5.4.1 Alignment . . . 133

5.4.2 Landmark Detection . . . 136

5.5 Model Implementation . . . 137

5.5.1 Early Approaches . . . 137

5.5.1.1 Neural Network Architecture . . . 137

5.5.2 DS13K . . . 137

5.5.3 DeepUAge . . . 142

5.5.4 Vec2UAge . . . 146

5.5.4.1 Dataset Curation . . . 146

5.5.4.2 Dataset Pre-Processing . . . 148

5.5.4.3 Dataset Split . . . 151

5.5.4.4 Dataset Augmentation . . . 151

5.5.4.5 Network Architecture . . . 152

5.5.4.6 Hyper-Parameters . . . 152

5.5.4.7 Improving Traditional Hash Sets with Facial Embedding Vectors . . . 153

**6 Results 155**

6.1 Facial Age Estimation Model Evaluation Platform . . . 155

6.1.1 Dataset Experiment Summary . . . 155

6.1.2 Entire Age Range Estimation . . . 156

6.1.3 Influence of Gender on Estimation . . . 158

6.1.4 Age Range Analysis . . . 159

6.2 Underage Age Estimation Improvement . . . 161

6.2.1 Dataset . . . 162

6.2.2 Underage Range Baseline Estimation . . . 162

6.2.3 DS13K Evaluation and Comparison . . . 163

6.2.4 Comparison with Ensemble Learning Techniques. . . . 164

6.3 VisAGe Dataset Generation Process . . . 168

6.3.1 Voting Statistics . . . 168

6.3.2 VisAGe Dataset Characteristics . . . 168

6.3.2.1 Dataset Sample and Distribution . . . 170

6.3.2.2 Physiological Predicted Variables . . . 173

6.3.2.3 Quality of Images . . . 173

6.3.2.4 Makeup and Accessories . . . 175

6.3.2.5 Facial Composition of the Subject . . . 177

6.3.3 Voting Human Effort . . . 182

6.4 Age Estimation Influencing Factors . . . 183

6.4.1 Microsoft Azure . . . 183

6.4.1.1 Strong PCC Distribution per Age with Error Difference Greater Than or Equal to 0 . . . 183

6.4.1.2 Error Distribution . . . 184

6.4.1.3 Strong PCC Distribution per Age With $\mathrm{Er_d}$ Less Than 5 185

6.4.1.4 Strong PCC Distribution per Age With $\mathrm{Er_d}$ Greater Than 5 . . . 185

6.4.1.5 Mild Correlations . . . 193

6.4.2 Amazon Rekognition (AWS) . . . 194

6.4.2.1 Error Distribution . . . 194

6.4.2.2 Strong PCC Distribution per Age with $\mathrm{Er_d}$ greater than 5 194

6.5 DeepUAge . . . 196

6.5.1 Evaluation . . . 196

6.5.1.1 DCA and Other Pre-Processing Techniques . . . 196

6.5.1.2 DeepUAge Performance . . . 197

6.5.1.3 Comparison of DeepUAge and Other Estimation Techniques . . . 198

6.6 Vec2UAge . . . 200

6.6.1 Evaluation of the Experiments . . . 200

6.6.2 D1 - Evaluation of the MAE Distribution with a Fixed Initial Learning Rate . . . 201

6.6.3 E1 - Evaluation of the MAE Distribution with an Initial LR Finder203

6.6.4 Details of Winning Optimisation Approaches . . . 204

6.6.5 Running Times and Convergence . . . . . . . . . . . . . . . . . . . 208
6.7 Summary of Model Results . . . . . . . . . . . . . . . . . . . . . . . . . . 210

**7 Conclusions and Future Work** **211**
7.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 211
7.2 Conclusions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 214
7.3 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 215

**Appendices** **239**

**A Ethical Approvals and Exemptions** **240**
A.1 LS-17-74-Anda-Scanlon . . . . . . . . . . . . . . . . . . . . . . . . . . . 241
A.2 LS-E-20-135-AndaBasabe-Scanlon . . . . . . . . . . . . . . . . . . . . . . 243

**B VisAGe** **244**
B.1 Audit Mechanism Document . . . . . . . . . . . . . . . . . . . . . . . . . 245
B.2 Dataset Collection System User Manual . . . . . . . . . . . . . . . . . . 246

# LIST OF FIGURES

2.1 CFFTPM Phases . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

2.2 Reports by Electronic Service Providers . . . . . . . . . . . . . . . . . . . 23

2.3 Pipeline of Designed Ensembles to Detect Child Sexual Exploitation Material . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

2.4 Binary Classification Structure . . . . . . . . . . . . . . . . . . . . . . . . 32

2.5 Regression Structure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32

2.6 Interest over Time of Machine Learning/Artificial Intelligence/Deep Learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

2.7 VGG16 Macro-Architecture . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.1 1000 Facial Landmarks Captured using Face++ Application Programming Interface . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 49

3.2 Distribution of Male and Female Subjects by Age - FG-NET . . . . . . 64

3.3 Distribution of Male and Female Subjects by Age and Ethnicity - MEDS 65

3.4 Ethnicity Distribution - Feret . . . . . . . . . . . . . . . . . . . . . . . . . 66

3.5 Distribution of Male and Female Subjects by Age Range - Adience . . 67

3.6 Demographic Distribution - UTKFace . . . . . . . . . . . . . . . . . . . 68

3.7 Distribution of Female Subjects by Age - Selfie-FV . . . . . . . . . . . 69

3.8 Demographic Distribution - MORPH . . . . . . . . . . . . . . . . . . . 71

3.9 Distribution of Age - IMDB-WIKI . . . . . . . . . . . . . . . . . . . . . . 72

3.10 Distribution of Age and Gender - YFCC100M . . . . . . . . . . . . . . 73

4.1 Minimum Mean Absolute Error per Year . . . . . . . . . . . . . . . . . 81

4.2 Cloud-Base Funding Timeline and Access to Resources. . . . . . . . . . 82

4.3 Curated Images per Age per Gender . . . . . . . . . . . . . . . . . . . . 84

4.4 Design Pattern UML Class Diagrams . . . . . . . . . . . . . . . . . . . . 86

4.5 Database Design . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
4.6 Dataset Generator System Architecture . . . . . . . . . . . . . . . . . . 88
4.7 Service Database Design . . . . . . . . . . . . . . . . . . . . . . . . . 90
4.8 VisAGe Data Collection System Architecture. . . . . . . . . . . . . . . . 94
4.9 UML Activity Diagram of VisAGe GUI . . . . . . . . . . . . . . . . . . 97
4.10 VisAGe Database Diagram . . . . . . . . . . . . . . . . . . . . . . . . 98
4.11 Facial Detection Approaches . . . . . . . . . . . . . . . . . . . . . . . 100
4.12 Loomis Face Proportion Approach Sketch . . . . . . . . . . . . . . . . 101
4.13 Dlib Contour Artistic Approach Process . . . . . . . . . . . . . . . . . 102
4.14 DS13K Average Faces between 16 to 17 Year Old's. . . . . . . . . . . . 104
4.15 VisAGe Dataset Average Image . . . . . . . . . . . . . . . . . . . . . 106
4.16 ResNet50 Pre-Trained on ImageNet with 20 Outputs in the Fully-Connected Softmax Layer . . . . . . . . . . . . . . . . . . . . . . . . 109
4.17 Histogram of Age Count Distribution: VisAGe, Selfie-FV and Combined . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
4.18 Facial Image Augmentation Techniques . . . . . . . . . . . . . . . . . 113
4.19 Vec2UAge - Proposed Neural Network . . . . . . . . . . . . . . . . . 114

5.1 Implementation of MongoDB Collections . . . . . . . . . . . . . . . . . 120
5.2 Dataset Generator Software . . . . . . . . . . . . . . . . . . . . . . . 121
5.3 Implementation of MongoDB Service Collections . . . . . . . . . . . . 125
5.4 Plotly Graph . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126
5.5 Contour Reconstruction from Dlib . . . . . . . . . . . . . . . . . . . . 136
5.6 Architecture of Multi-Class Prototype Model . . . . . . . . . . . . . . . 138
5.7 DS13K Caffe Graph . . . . . . . . . . . . . . . . . . . . . . . . . . . . 141
5.8 DeepUAge Folder Structure . . . . . . . . . . . . . . . . . . . . . . . 144
5.9 VisAGe Image Structure before Merge. . . . . . . . . . . . . . . . . . 148

6.1 Average Estimated Age Compared with Actual Age across Entire Dataset. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 157
6.2 Average Estimated Age Compared with Actual Age . . . . . . . . . . . 159
6.3 Mean Average Error Rate . . . . . . . . . . . . . . . . . . . . . . . . . 160
6.4 Mean Average Error Rate for Each System in Different Age Ranges. . 161
6.5 Average Estimated Age from Each Service Compared with Actual Age. 162
6.6 Mean Absolute Error per Age by Service. . . . . . . . . . . . . . . . . 163
6.7 Age Estimation Accuracy per Age Group. . . . . . . . . . . . . . . . . 166

6.8 Number of Positively and Negatively Resolved Votes per Age Bracket 169
6.9 Number of Deleted Photos per Age of Subject. . . . . . . . . . . . . . . . . 170
6.10 VisAGe Dataset Age Distribution by Gender. . . . . . . . . . . . . . . . . 171
6.11 Image Average VisAGe. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 172
6.12 Mean Height and Width per Age. . . . . . . . . . . . . . . . . . . . . . . . 174
6.13 Makeup and Accessories per Age Categories . . . . . . . . . . . . . . . . 176
6.14 Predicted Male Facial Hair Counts per Age. . . . . . . . . . . . . . . . . 177
6.15 General Distribution of Emotions. . . . . . . . . . . . . . . . . . . . . . . . 178
6.16 Mean Smile Intensity per Age per Gender. . . . . . . . . . . . . . . . . . 179
6.17 Mean Head Pose Angle per Age. . . . . . . . . . . . . . . . . . . . . . . . . 180
6.18 Counts of Hair Colour per Age. . . . . . . . . . . . . . . . . . . . . . . . . . 181
6.19 Total of Voting Hours per Date per User . . . . . . . . . . . . . . . . . . . 182
6.20 Azure: Distribution of $Er_d$ per age. . . . . . . . . . . . . . . . . . . . . . . 184
6.21 Azure: Box-plot of PCC Distribution per Age, where $Er_d$>5. . . . . . . 186
6.22 Azure: Strong Correlations between Attributes and $Er_d > 5$ for Ages 1 and 2. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 187
6.23 Azure: Strong Correlations between Attributes and $Er_d > 5$ for Ages 4 and 5. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 188
6.24 Azure: Strong Correlations between Attributes and $Er_d > 5$ for Ages 6 and 7. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 189
6.25 Azure: Strong Correlations between Attributes and $Er_d > 5$ for Ages 9 and 10. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 190
6.26 Azure: Correlations per Age with $Er_d > 5$. . . . . . . . . . . . . . . . . . 192
6.27 AWS: Distribution of $Er_d$. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 194
6.28 AWS: Correlations per Age with $Er_d > 5$. . . . . . . . . . . . . . . . . . . 195
6.29 DeepUAge - Real Age vs Predicted Age . . . . . . . . . . . . . . . . . . . 197
6.30 Average Difference per Age per Facial Age Predictor. . . . . . . . . . . 199
6.31 Mean Absolute Error Distribution per Optimiser for Training/Validation/Test (Using a Fixed Initial Learning Rate) . . . . . . . . . . . . . 202
6.32 Mean Absolute Error Distribution per Optimiser for Training/Validation/Test (Using LR Finder) . . . . . . . . . . . . . . . . . . . . . . . . 204
6.33 Performance per Age for the Top 3 Best Performers . . . . . . . . . . . 207
6.34 Average Runtime per Optimisation Algorithms for Experiments *D1* and *E1* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 209

# LIST OF TABLES

2.1 Cybercrime Classification . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

2.2 Actions and Phase for the COMMONWEALTH vs. CHARLES HINDS, JR Case. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 15

3.1 Distribution of Image Quality - MORPH . . . . . . . . . . . . . . . . . . . 70

3.2 Best Performing Offline Approach per Year . . . . . . . . . . . . . . . . . . 77

3.3 Summary of Facial Datasets for Age Estimation. . . . . . . . . . . . . . 77

4.1 Cloud-Based Facial Age Estimation Services . . . . . . . . . . . . . . . . . 82

4.2 CC Licenses used by Flickr . . . . . . . . . . . . . . . . . . . . . . . . . . . 92

4.3 Variables with the Corresponding Weights and Descriptions for Image Selection. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 96

4.4 VisAGe Dataset Demographics . . . . . . . . . . . . . . . . . . . . . . . . 106

4.5 Microsoft Azure Cognitive Services Face API Attributes . . . . . . . . 116

4.6 Amazon AWS Rekognition Attributes . . . . . . . . . . . . . . . . . . . . 117

5.1 Dataset Generator Executable Arguments . . . . . . . . . . . . . . . . . . 122

5.2 Service Evaluation Executable Parameters . . . . . . . . . . . . . . . . . 123

5.3 Description of VisAGe Collections . . . . . . . . . . . . . . . . . . . . . . 131

5.4 VisAGe Metadata Format . . . . . . . . . . . . . . . . . . . . . . . . . . . 132

5.5 VisAGe URL Routes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133

5.6 Cosine Similarity . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 152

6.1 Mean Absolute Error per Service. . . . . . . . . . . . . . . . . . . . . . . . 156

6.2 Mean Absolute Error per Gender per Service. . . . . . . . . . . . . . . . 158

6.3 Performance per Age Range. . . . . . . . . . . . . . . . . . . . . . . . . . 161

6.4 Mean Absolute Error for Underage Images per Service. . . . . . . . . . 163

6.5 Accuracy per Group per Service. . . . . . . . . . . . . . . . . . . . . . . . . 164
6.6 Mean Absolute Error Rates for the 0-25 Age Range. . . . . . . . . . . . 165
6.7 Several regression approaches per age range . . . . . . . . . . . . . . . . . 166
6.8 Statistics of Image Dimensions . . . . . . . . . . . . . . . . . . . . . . . . . . . 173
6.9 Microsoft Azure Face Application Programming Interface Quality Attributes. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 174
6.10 Statistics of Head Pose. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 180
6.11 Azure facial attributes: Strong (Black) and Mild (Grey) Influencing Factors with $\mathrm{Er_d} > 5$. . . . . . . . . . . . . . . . . . . . . . . . . . . 191
6.12 Results of Different Pre-Processing Techniques . . . . . . . . . . . . . . 196
6.13 Evaluation of Facial Age Estimators for Underage Subjects . . . . . . . 198
6.14 Experiment D1: Statistics of MAE in Training/Validation/Test (Fixed LR of `1e−4` for ADAGRAD, ADAM and SGD; and `1e−5` for SWA) . 202
6.15 Experiment E1: Statistics of Mean Absolute Error in Training/Validation/Test (LR Finder Executed) . . . . . . . . . . . . . . . . . . . . . . . . 204
6.16 Experiment D1: Validation Mean Absolute Error with Stochastic Weight Averaging Optimiser and Fixed LR of `1e−5`. . . . . . . . . . . 205
6.17 Experiment E1: Validation Mean Absolute Error with Stochastic Gradient Descent Optimiser and LR Finder. . . . . . . . . . . . . . . . . 206
6.18 Summary of Created Models . . . . . . . . . . . . . . . . . . . . . . . . . . 210

# LIST OF ALGORITHMS

1 Robust-Hashing - PhotoDNA . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
2 Dlib Contour Artistic Approach . . . . . . . . . . . . . . . . . . . . . . . . . . . . 136
3 Facial Embedding Image Recognition Approach - FaceNet . . . . . . . . . . 154

# LIST OF CODE SNIPPETS

1 Database Superclass . . . . . . . . . . 119
2 Morph Methods . . . . . . . . . . 119
3 FGNET Record Insertion Example . . . . . . . . . . 120
4 Dataset Generator Execution Command . . . . . . . . . . 122
5 Evaluation Platform Execution Command . . . . . . . . . . 122
6 Service Superclass . . . . . . . . . . 123
7 Microsoft Azure Methods . . . . . . . . . . 124
8 DEX Record Insertion Example . . . . . . . . . . 125
9 Regular Expression for Age Text Search . . . . . . . . . . 128
10 Background Scripts Method Class . . . . . . . . . . 129
11 Microsoft Azure Facial Attribute Prediction Response . . . . . . . . . . 130
12 DCA - Align Code . . . . . . . . . . 135
13 DS13K Architecture . . . . . . . . . . 139
14 DeepUAge - Dataset Split . . . . . . . . . . 143
15 DeepUAge - Dataset Augmentation . . . . . . . . . . 145
16 Keras ResNet50 Model Creation . . . . . . . . . . 146
17 File Normalisation Script . . . . . . . . . . 147
18 Dlib Facial Detection Types . . . . . . . . . . 149
19 FaceNet Vector Extraction . . . . . . . . . . 150
20 Vec2UAge - Dataset Split . . . . . . . . . . 151
21 Neptune Configuration . . . . . . . . . . 153

# LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| **ADAGRAD** | Adaptive Gradient | 46 |
| **ADAM** | Adaptive Moment Estimation | 46 |
| **AI** | Artificial Intelligence | 2 |
| **ANN** | Artificial Neural Network | 46 |
| **API** | Application Programming Interface | 49 |
| **AWS** | Amazon Web Services | 81 |
| **BIF** | Biologically Inspired Features | 54 |
| **CAM** | Child Abuse Material | 3 |
| **CC** | Creative Commons | 31 |
| **CCTV** | Closed-circuit Television | 16 |
| **CF** | Computer Forensics | 12 |
| **CFAA** | Computer Fraud And Abuse Act | 9 |
| **CFFTPM** | Cyber Forensic Field Triage Process Model | 13 |
| **CIA** | Confidentiality, Integrity And Availability | 11 |
| **CL** | Curriculum Learning | 34 |
| **CLI** | Command Line Interface | 125 |
| **CNN** | Convolutional Neural Network | 39 |
| **CODIS** | Combined DNA Index System | 63 |
| **CP** | Child Pornography | 24 |
| **CPU** | Central Processing Unit | 208 |
| **CRC** | Computer Related Crimes | 2 |
| **CSAM** | Child Sexual Abuse Material | 11 |
| **CSEM** | Child Sexual Exploitation Material | 1 |
| **CSP** | Cloud Service Provider | 51 |
| **CSV** | Comma-separated Values | 131 |
| **DCNN** | Deep Convolutional Neural Network | 56 |
| **DDoS** | Distributed Denial-of-service | 7 |
| **DEX** | Deep EXpectation | 39 |
| **DF** | Digital Forensics | 4 |
| **DFI** | Digital Forensic Investigator | 20 |
| **DFL** | Digital Forensic Laboratory | 19 |
| **DL** | Deep Learning | 1 |
| **DNN** | Deep Neural Network | 45 |
| **DoS** | Denial-of-service | 10 |

| | | |
|---|---|---|
| **ESP** | Electronic Service Providers | 22 |
| **FBI** | Federal Bureau Of Investigation | 9 |
| **FC** | Fully-connected | 43 |
| **FCE** | Face Colour Extraction | 60 |
| **FTK** | Forensic Toolkit | 18 |
| **FTP** | File Transfer Protocol | 17 |
| **GAN** | Generative Adversarial Network | 34 |
| **GCHQ** | Government Communications Headquarters | 18 |
| **GCP** | Google Cloud Platform | 81 |
| **GDPR** | General Data Protection Regulation | 40 |
| **GPU** | Graphics Processing Unit | 39 |
| **GUI** | Graphical User Interface | 85 |
| **HBFT** | Hierarchical Bloom Filter Tree | 28 |
| **HDD** | Hard Disk Drive | 9 |
| **HOG** | Histogram-oriented Gradient | 103 |
| **HRN** | Human Rights Now | 25 |
| **HSV** | Hue, Saturation, Value | 60 |
| **ICMP** | Internet Control Message Protocol | 17 |
| **ICT** | Information Communications Technology | 7 |
| **IIOC** | Indecent Images Of Children | 24 |
| **IoT** | Internet Of Things | 19 |
| **IP** | Intellectual Property | 10 |
| **IQR** | Interquartile Range | 185 |
| **IRC** | Internet Relay Chat | 17 |
| **JSON** | JavaScript Object Notation | 56 |
| **KD** | Knowledge Distillation | 38 |
| **KNN** | K-nearest Neighbours | 57 |
| **KPI** | Key Performance Indicator | 8 |
| **KVM** | Kernel-based Virtual Machine | 22 |
| **LEA** | Law Enforcement Agency | 1 |
| **LR** | Learning Rate | 45 |
| **MAD** | Mean Absolute Difference | 37 |
| **MAE** | Mean Absolute Error | 5 |
| **MEDS** | Multiple Encounter Dataset | 65 |
| **ML** | Machine Learning | 2 |
| **MLaaS** | Machine Learning As A Service | 56 |
| **MLP** | Multilayer Perceptron | 58 |
| **MOBA** | Multiplayer Online Battle Arena | 7 |
| **MSE** | Mean Squared Error | 37 |
| **MTCNN** | Multitask Cascaded Convolutional Network | 50 |
| **MVC** | Model-view-controller | 85 |
| **NC** | Netcat | 17 |
| **NCCS** | National Computer Crime Squad | 11 |
| **NCMEC** | National Center For Missing And Exploited Children | 17 |
| **NGO** | Non-governmental Organisation | 25 |
| **NSRL** | National Software Reference Library | 27 |
| **OS** | Operating System | 3 |

| | | |
|---|---|---|
| **P2P** | Peer-to-peer | 16 |
| **PCA** | Principal Component Analysis | 48 |
| **PCC** | Pearson Correlation Coefficient | 116 |
| **PDF** | Portable Document Format | 92 |
| **PI** | Principal Investigator | 92 |
| **PSDN** | Public Switched Data Network | 10 |
| **PSN** | Public Switched Network | 9 |
| **PSTN** | Public Switched Telephone Network | 9 |
| **RAM** | Random Access Memory | 110 |
| **ReLU** | Rectified Linear Unit | 43 |
| **RMSE** | Root Mean Squared Error | 37 |
| **SEIC** | Sexually Exploitative Imagery Of Children | 23 |
| **SGD** | Stochastic Gradient Descent | 46 |
| **SIS** | Schengen Information System | 30 |
| **SSD** | Solid State Drive | 19 |
| **SSH** | Secure Shell | 92 |
| **SSKD** | Simple Skin Detection | 60 |
| **SVM** | Support Vector Machine | 18 |
| **SWA** | Stochastic Weight Averaging | 46 |
| **SWAG** | SWA-Gaussian | 216 |
| **UCD** | University College Dublin | 82 |
| **UML** | Unified Modeling Language | 85 |
| **URL** | Uniform Resource Locator | 94 |
| **VisAGe** | Visual Age And Gender | xxi |
| **VM** | Virtual Machine | 18 |
| **VPN** | Virtual Private Network | 8 |

# ABSTRACT

We are living in a digital era where most transactions are contact-less, social media platforms are commonplace and a part of our daily life is recorded either in a permissive or surreptitious manner. Whether we are present in an online meeting, daily social media feed, a peer-connected calendar, a live gaming or video stream, hundreds of bytes of our information are sent through a network to a server. The exponential growth of storage is also enabling thousands of multimedia content to be stored locally on digital devices but at the same time challenging digital investigations that are hampered by the accumulation of such devices that were stored in a forensic laboratory awaiting to be processed by an expert in a timely manner. The size and amount of information that requires analysis is increasing, leading to an ungovernable digital forensic backlog. Smartphone users are able to produce original content such as audio, images and videos, and thanks to the internet, are able to broadcast data worldwide in a matter of seconds. Digital forensic practitioners have become overwhelmed by the amount of data that they encounter and are requiring the implementation of artificial intelligence as tools and techniques to aid investigations, to discover, gather and analyse records swiftly.

To address the digital forensic backlog, the creation of age estimation models to assist digital forensic investigations has been proposed. Although some models perform well for the entire age range, in certain age ranges such as the underage group, the models perform wholly inadequate. Influencing factors on underage age estimation have been evaluated and it has been determined that certain elements have strong, mild or weak correlations with the machine-predicted performance. These considerations are key on the curation of datasets and will yield better results on future trained models.

The largest underage dataset with age and gender labels has been collected and several models have been experimented with different image pre-processing techniques, neural network architectures, etc. Hyper-parameter optimisation was introduced and the best score for facial age estimation was obtained. The scores were evaluated with a chosen test dataset that contains faces that can be spotted by well-known face detectors such as Viola Jones. A novel facial embedding approach was proposed and a distribution evaluation metric was introduced instead of a single value. The performance achieved surpasses the state-of-the-art facial age detectors for subjects under the age of 25.

# STATEMENT OF ORIGINAL AUTHORSHIP

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the title page and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

# COLLABORATIONS

Several collaborations were achieved throughout the path of the research that assisted the production of this thesis. Firstly, research within the UCD Forensics & Security Research Group was prevalent and contributed widely to the development of this work.

Second, the label annotation of the Visual Age and Gender (VisAGe) dataset discussed in Section 4.3.2 was assisted in a collaboration between the UCD Forensics & Security Research Group and MITRE. Both teams provided resources and support to aid the creation of the largest human verified and curated facial age dataset for underage subjects.

Lastly, work on facial age estimation through facial vector embeddings was promoted by Edward Dixon from Intel Corporation, and the collaborative research is discussed in Section 3.2.5. The approach developed resulted in the achievement of a machine learning model for facial age prediction of minors that surpasses the state-of-the-art models.

# ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and encouragement of my family. Their unconditional support throughout the years. I am also thankful to my supervisors, Dr. Mark Scanlon and Dr. Nhien-An Le-Khac. In particular, I would like to thank Mark, for guiding me since the first day I stepped into UCD and having the time to guide me not only in the academic but personal aspects. With almost every weekly meetings available while mentoring and nurturing my academic interests.

I would also like to thank the UCD School of Computer Science for funding my research entirely. Without their help, It would not have been possible to study at one of the top Universities in Ireland with all the equipment and physical space provided, being suitable for a pleasant working environment.

My gratitude to the UCD IT Services for providing access to the Sonic Server for which a great number of my experiments were conducted. Without the server I had to create external tedious applications hoping to get funding each time.

Amazon Web Services and Microsoft Azure contributed generously to this research with grants, funding the execution of experiments that required GPU and high end processors.

Many thanks to my friends Jake, Dave and my wife for being part of our Saturday MTG Group. While doing activities other than research that helped strengthening our social skills and mental health in times of a pandemic.

# DEDICATION

To the memory of my father, Eduardo Augusto Anda who would have been proud of me following his steps into academia. His inspiration has led me to strive and achieve goals that would have been unattainable. To my mother, Graciela, who has supported and provided affection and love through my entire life. To my beloved Grandmother, who has showered me with support not only during the process but whenever needed. And to my siblings who have always been there to offer their help. I also want to dedicate this thesis to my baby son who has been encouraging me day to day to be strong, organised and caring while continuing my voyage towards completing my PhD. A special dedication to my wife who has been with me in this adventure, while supporting and loving me, nourishing and protecting our son.

# LIST OF PUBLICATIONS

Journals

- Anda, F., Dixon, E., Bou-Harb, E., Le-Khac, N. A., & Scanlon, M. (2021). Vec2UAge: Enhancing Underage Age Estimation Performance through Facial Embeddings. Forensic Science International: Digital Investigation, Volume 36, Supplement, 2021, 301119, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2021.301119.

- Anda, F., Le-Khac, N. A., & Scanlon, M. (2020). DeepUAge: Improving Underage Age Estimation Accuracy to Aid CSEM Investigation. Forensic Science International: Digital Investigation, Volume 32, Supplement, 2020, 300921, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2020.300921.

- Anda, F., Lillis, D., Kanta, A., Becker, B. A., Bou-Harb, E., Le-Khac, N. A., & Scanlon, M. (2019). Improving the accuracy of automated facial age estimation to aid CSEM investigations. Digital Investigation, Volume 28, Supplement, 2019, Page S142, ISSN 1742-2876, https://doi.org/10.1016/j.diin.2019.01.024

Conference Papers

- Anda, F., Becker, B. A., Lillis, D., Le-Khac, N. A., & Scanlon, M. (2020, June). Assessing the Influencing Factors on the Accuracy of Underage Facial Age Estimation. In 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1-8). IEEE.

- Du, X., Hargreaves, C., Sheppard, J., Anda, F., Sayakkara, A., Le-Khac, N. A., & Scanlon, M. (2020, August). SoK: exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation. In Proceedings of the 15th International Conference on Availability, Reliability and Security (pp. 1-10).

- Anda, F., Lillis, D., Kanta, A., Becker, B. A., Bou-Harb, E., Le-Khac, N. A., & Scanlon, M. (2019, August). Improving borderline adulthood facial age estimation through ensemble learning. In Proceedings of the 14th International Conference on Availability, Reliability and Security (pp. 1-8).

- Anda, F., Lillis, D., Le-Khac, N. A., & Scanlon, M. (2018, May). Evaluating automated facial age estimation techniques for digital forensics. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 129-139). IEEE.

Posters

- Anda, Felix & Le-Khac, Nhien-An & Scanlon, Mark. (2017). Automated Machine Learning-Based Digital Evidence Classification Techniques, 16th European Conference on Cyber Warfare and Security (ECCWS), Dublin, Ireland.

In Preparation

- Anda, F., Becker, B. A., Hall, C., Doyle, J. S., Lillis, D., Le-Khac, N. A., & Scanlon, M. (2020). VisAGe: Visual Age and Gender Dataset.

CHAPTER

ONE

# INTRODUCTION

## 1.1 Preamble

The central theme of this dissertation is based on cybercrime investigations and the application of deep learning (DL) to assist investigators in identifying missing children, victims, suspects and combating the accumulation of unprocessed digital devices (known as digital forensic backlog) seized in crime scenes.

The influence of digital data in our daily life has increased the number of apprehended devices in a crime scene. Scanning the surface of a disk for digital evidence has long been a time-consuming task for forensic investigators. Law enforcement agencies (LEAs) struggle to obtain information with probative value due to the lack of time and experts required to process such devices. Once the evidence is collected, it is stored in a digital forensic laboratory awaiting to be analysed. But due to the mentioned lack of resources, there is an increment in the backlog. In order to process the considerable amount of data seized and processed in a typical case, a time-consuming, highly-skilled digital forensic analysis must be conducted. Furthermore, the seized devices must often be processed immediately due to the urgent need of evidence to progress an investigation that could be a matter of life or death. [217].

Human soft biometric traits, such as age and gender, have been used by police officers and witnesses in their description of unidentified victims. In certain cases, the age of the victim can result in the determination of a crime categorisation. Production, distribution, and possession of child sexual exploitation material (CSEM) are considered illegal activities in most jurisdictions. The growth of CSEM has been influenced by its availability on the deep web and has resulted in a large expansion of cybercrime. During CSEM investigations, digital forensic investigators are exposed to such material and some involved personnel are negatively impacted due to such contact [261].

Several studies have revealed digital forensic backlogs in Ireland, were ranging from one to four years [108, 122, 225]. It has been reported that the average digital forensic backlog in laboratories around the globe was from six to twelve months in 2009 [40]. The Irish media has recently reported in early January 2020 that computer related crimes (CRC) are experiencing up to 5 years of delay due to the law enforcement backlog and cases have been held and ruled upon in court with potentially pertinent evidence sitting untouched in the evidence lockers. In an attempt to address the backlog issue, the Cybercrime Bureau in An Garda Síochána has been boosted to over 30 members in the past and recently, jumped to over 200 personnel. Nevertheless, it is unsustainable that LEAs are able to sufficiently increase the number of digital forensic expert personnel. Therefore, the use of artificial intelligence (AI) techniques to aid digital forensic examinations is proposed. The leveraging of branches of AI, such as machine learning (ML), DL, and computer vision to lessen the digital forensic backlog is a promising approach. However, such approach is at its earliest stage and has recently started to contribute to the digital forensic community.

With the introduction and growth of modern information technology, the evolution of evidences held in court has migrated from just the "traditional" physical evidence to the inclusion of digital evidence which is widely scattered in both local and cloud environments. The significant lack of resources and automation in the acquisition and/or analysis steps of the digital forensic process increments the digital forensic backlog. Per Scanlon, the lack of these resources will continuously influence the throughput of digital forensic laboratories and therefore, are likely to continue hindering investigators in the future [225]. To address the digital forensic backlog, we have focused solely on the facial age component which can be further implemented in a pipeline of CSEM detection systems (refer to Section 2.2) or as a component in cloud environments such as Hansken [244] or Katalyst[1].

While human capabilities to detect and identify multiple facets, such as age, gender, ethnicity and facial expressions, can be accomplished by a quick glance at a digital image, machines are required to be trained intensively in order to understand traits present in photographs. Facial recognition has been the main attraction of several products in these last couple of years and performance controversy has returned to the mainstream media due to the new government regulations on mask coverings in crowded places, impeding users to authenticate properly on certain devices as a result of occlusion. Facial recognition technology as unlocking/authentication security mechanism surpasses the traditional fingerprint authentication. China has used facial recognition technology across multiple applications, e.g., identification of ride hailing service driver and jaywalkers, pay with a smile, etc. In the USA, it has been used in

[1] `https://projectvic.org/katalyst/`

religious services to track worshippers, and in the UK, it has been used to stop shoplifters.

As with any type of technology, there are also drawbacks to facial recognition such as privacy concerns, violations of rights and personal freedom, and biased algorithms that do not perform as well for certain ages/gender/race, etc. Diversity of the data and the collection of large unbiased datasets are key to address this issue and is encouraged in our future work.

The appearance of ubiquitous object detectors will generate a considerable impact on the life of users and consumers. Migrating to a facial-recognition-based authentication factor can strengthen system security to prevent impersonation attacks. Facial security checks could impede card cloning, fraudulent exam takers, and identity theft. Clustering photos of a same subject is a functionality that has been used in several operating system (OS) photo album features. This feature could assist investigations and backlogs and has been evaluated in this study through the use of facial embedding vectors.

The use cases for accurate age estimation are not only limited to child abuse material (CAM) investigation but are useful across a range of crimes such as exploitation, trafficking, forced labour and abduction. Age estimation commodities are becoming more common in our milieu. Applications can be found in adult entertainment venue access, purchase of age-restricted goods such as alcohol, tobacco and lottery, and age targeted advertising. The aforementioned scenarios are just some examples of the variety of applications that can be achieved with "multi-layered DL technology for highly intelligent services" [258].

The accurate age estimation of a subject has always been a challenge for research across several fields. From counting the annual rings of wood growth to determine the age of a tree to measuring the skeletal maturity of a bone to obtain the age of a living being. Several methods have been proposed, from measurement-driven anthropomorphic analysis to the application of ML algorithms, and the accuracy is constantly improving. Furthermore, the demand to distinguish the marginally under-age from the slightly overage is a matter of study where existing methods have been proven not to be reliable enough to be able to perform the task.

## 1.2 Research Ethics

Prior to data collection, institutional ethical approval was obtained (University College Dublin Human Research Ethics Committee reference number LS-17-74-Anda-Scanlon). Refer to Appendix A.1 for more information. The dataset is subject to reinforced safeguards to prevent unauthorised distribution. The data controller analyses the applications and allows or denies access to researchers that have been registered adequately.

It was also necessary to declare that the use of the facial embedding dataset (Selfie-FV) was exempt from a full ethical review. This due to the data consumed being in the form of facial vectors which is further discussed in Section 3.3.4.1. The data used is completely anonymous and obtaining personal identifiable information from a vector is infeasible. Further information regarding the exemption may be consulted in the Appendix A.2.

## 1.3 Research Questions

The main aim of this research is to aid investigations by introducing a novel approach to identify victims and/or suspects with soft biometric cues such as age and gender in an automated manner. While contributing with robust forensically sound tools and techniques for digital forensics (DF) that can support the evidence that is presented in court in a timely manner, address the digital forensic backlog that has become a well know issues in LEAs throughout the globe.

The research questions are the following:

1. What are the influencing factors that can improve the performance of facial age prediction models?
2. How can DL be used to aid digital forensic investigators and lessen their exposure to sensitive data?
3. How can the design and implementation of an age-prediction-based DL model assist the digital forensics backlog?
4. What is the impact of integrating ML and DL techniques with digital forensic processing with regards to forensic soundness, court admissibility, and case throughput capabilities?

## 1.4 Contributions

1. Scientific
    - Evaluation of the current state of the art in facial age estimation.

- Comprehensive review of age estimation techniques, and the influencing factors on the performance.
- Experimental evaluation of current age estimation techniques both offline and cloud-based.
- Collection and dissemination of the largest curated underage facial age dataset with exact year and month of age and gender labels (VisAGe dataset).
- Implementation of a novel offline pre-processing technique (suitable for LEAs) for face landmark detection and hairline contour prediction of single frontal face images: contour artistic `dlib` approach.
- Design of a classification model based on a ResNet50 architecture with a mean absolute error (MAE) of 2.73 for under-age subjects (DeepUAge).
- Implementation of a novel regression model based on facial vector embeddings with a MAE distribution of 2.5 and a winning model of 2.3 (Vec2UAge)
- Significant improvement over individual cloud-based age estimators through the use of ensemble-based approaches for subjects under the age of 18 - comparable with expert human estimators.
- Modelling of a hyper-parameter optimisation approach that yielded a winning model of 2.5 (MAE).

2. Technical

   - The release of a standalone client-based tool designed to generate unique, uniformly distributed random images by age and gender from several facial image datasets (such as FG-NET, FERET, IMDB-WIKI, MEDS, YFCC100M).
   - Creation of an online collaborative voting-based dataset collection framework with age and gender annotation capacities (VisAGe data collection system).

## 1.5 Layout of this Thesis

- Chapter 2 introduces a comprehensive literature review of the background and related work of this research.

- Chapter 3 describes the related work that has been done in the fields of facial age datasets, data bias, human and machine estimations, influencing factors of facial ageing, facial vectors, the digital forensic backlog and adult content detection approaches.

- Chapter 4 outlines the methodologies for evaluating facial age estimation, improving the borderline adulthood age estimation, VisAGe (a framework for facial image collection and age and gender annotations), Creation of models and the assessment of the influencing factors on the accuracy of facial age estimation.

- Chapter 5 demonstrates the implementation of the dataset generator, the online collaborative voting-based dataset collection framework with age and gender annotation capacities, a novel pre-processing approach and the age estimation models.

- Chapter 6 describes the results of the evaluation of age estimation models, and our own developed models, the outcome of VisAGe is also presented. And finally, age estimation performance influencing factors are listed.

- Chapter 7 presents the conclusion and future work.

CHAPTER

**TWO**

# BACKGROUND RESEARCH

## 2.1 Cybercrime and Digital Forensic Investigations

### 2.1.1 Cybercrime

#### 2.1.1.1 Introduction

Criminal offences are illegal actions penalised accordingly by the relevant jurisdiction[1]. In Ireland, crimes committed with information communications technology (ICT) tools are a persistent and prevalent problem as it is throughout the globe. The term cybercrime is usually referred to as a crime committed in cyberspace against property, individuals or organisations with the use of digital devices or other forms of ICT. Property may refer to bank and utility accounts, electronic devices, home assistants that control lights, door bells, learning thermostats, smart plugs, smart TVs, etc. Cyberstalking, cyberbullying, production, possession and dissemination of CSEM, and trafficking are categorised as cybercrime against individuals. Organisations and governments are prone to become victims of cyber-terrorism, distributed denial-of-service (DDoS) attacks, hackitvism for a socially or politically motivated reason, web defacement, etc. Individuals are also susceptible to DDoS attacks particularly in multiplayer online battle arena (MOBA) games.

Hunton defined cybercrime as undesirable behaviour and other illicit activities that require the use of networked telecommunications [114]. Cyber-dependant crime by definition, involves the use of technology such as the internet, digital devices, networks and other means of communication. It is one of the fastest growing criminal activities across the world and has yielded yearly economical damages of over €300 billion [214]. These investigations entail the collection of digital evidence from sev-

[1] Official power to make legal decisions and judgements

eral sources [18]. Furthermore, involving the process of investigating and applying digital forensic techniques (explained in Section 2.1.2) to preserve, analyse and recover digital forensic data in an internet-based or local crime scene and further identify perpetrators and their true motivations. Criminal or *mens rea* (illicit intent which leads to certain criminal behaviour) is a key factor of criminology [114] and will be prevalent in digital-oriented investigations throughout the world. Investigations on cybercrimes can be both proactive and reactive. For the former, in response to intelligent systems such as anomaly detection, pattern recognition or other systems with AI capabilities that can detect potential threats and attempts to vulnerate systems. For the latter, after the crime has been identified, the respective notification to the relevant authorities.

For this thesis, an emphasise on cybercrimes against individuals will be accomplished; these type of crimes may not affect the global economy as much as other type of cyber-dependant crimes but are more harmful to humans, specifically children and should become a top priority for LEAs. In this section, challenges and classifications per authors are presented.

#### 2.1.1.2 Challenges

It has been argued that cyber-dependant investigations are still in their early stages of infancy [114], and much can be learned from the development of other established sciences. Jewkes and Andrews reported in 2005, that the police service tends to respond to innovation with an increased delay [128]. The aforementioned report established that priorities and performance of LEAs in England & Wales were found to be measured by government-set key performance indicators (KPIs) that excluded CSEM [128]; thus hampering top priority investigations for crimes against children in the internet.

Bequai [24] has reported that nearly 90% of financial industries lack the capability to investigate cybercrimes. One of the main obstacles in cybercrime is the transnational nature, where the offence can transcend worldwide with lack of international cooperation to investigate the case, resulting in high costs and time-consuming research.

Specialised police cyber-crime units face technical challenges due to the lack of training and experience and limited abilities to conduct cybercrime investigations [152], whereas cybercriminals are usually technologically-aware and constantly developing and adapting to new tools to allow them to be one step ahead of LEAs [189].

Anonymity and attribution are two key factors that complicate investigations. The former is feasible due to a plethora of techniques such as proxies, onion servers, mixed networks, virtual private networks (VPNs), etc. The latter is slightly more complex; although a system can record logs with nonrepudiation, the integrity of a transaction could be compromised by savvy malicious users with the use of techniques such as SQL injection, social engineering, or ultimately, employing the trojan horse defence in court so the defendant can claim he did not commit the *actus reus* [33]. Furthermore, a person can impersonate another through elaborate spoofing schemes with ease [123].

The ever-increasing storage capacity of hard disk drive (HDD) has presented a huge challenge for cybercrime investigations, contributing to the increase of the digital forensic backlog (Outlined in Section 2.1.2.7). Data with probative value may be hidden through steganography or encryption [123], thus increasing the processing times of an investigation.

#### 2.1.1.3 Classification of Cybercrime

Cybercrimes were vaguely prosecuted as mail and wire fraud until the appearance of the first federal computer fraud law: Computer Fraud and Abuse Act (CFAA) in 1984, designed to criminalised unauthorised access to computers. Ever since, the CFAA has faced several modifications [135], being the latest update in 2008 and recently discussed in the Obama administration back in 2015.

The CFAA included seven types of criminal activity [124]:

- Cyber Espionage,
- Obtaining Information by Unauthorized Computer Access,
- Government Computer Trespassing,
- Computer Fraud,
- Damaging a computer
- Password trafficking, and
- Threats and extorsion.

Cybercrime classification has evolved throughout the years and suffered several alterations due to the constant technological development. In the late 90's, the Federal Bureau of Investigation (FBI) introduced a classification of cyber-dependant crimes in seven segments which struggled to contemplate most of the crimes for that time. Furthermore, it was observed that CSEM which is a subject discussed in this dissertation in Section 2.2, is vaguely classified in the "Others" region. The first classification included intrusions of the public switched network (PSN), which refers to carrier networks that provide circuit switching for public users. It is a term applied to public switched telephone network (PSTN) but can also be referred to public switched data network (PSDN). The following divisions are major computer network intrusions, network integrity and privacy violations, industrial espionage, pirated computer software and “others”, which is a class that represents other crimes where the computer is a major factor in committing the criminal offence.

Later in 1998, the UK Audit Commission developed a classification methodology of 9 items [16]. It is clear that through the years, additional cybercrimes were identified. According to Dowland, offence involving pornographic material, were not reported until the year 2001 [65]. In that year, a total of 625 reported incidents of computer crime and abuse in the UK were documented. Pornography was the second most reported incident after viruses with 193 instances equating to a 30% of the total.

Wall has defined several approaches to cybercrime classification [250, 251, 252]. His first approach in 2001, categorised in: cyber-trespass which links to crossing boundaries into other people’s property and/or causing damage (hacking, defacement, viruses), cyber-deceptions & thefts that refer to stealing money and intellectual property (IP) theft in general, cyber-pornography refering to breaching laws on obscenity and decency, and cyber-violence which refers to psychological harm through hate speech, stalking, etc. His second approach which was documented in 2007, divides the categories in 3 broad groups: computer integrity, computer assisted and computer content crimes; it can be noticed that now there is a clear classification for CSEM through content. Finally in 2015, Wall defines similarly 3 groups namely crimes “against”, “using” and “in” the machines. The crimes in the machines category is similar to the computer content crimes previously defined in 2007.

In 2001, Furnell designed a high level of categorisation for cybercrimes which were based on the UK Audit Commission [83]. They were classified into two simple categories: “computer-assisted” and “computer-focused” crimes. A year later, James and Nordby also maintained a dual classification system consisting in computer as an instrument in criminal activity and computer as a target of criminal activity. The former contains activities such as child pornography and solicitation, stalking and harassment, fraud, software piracy, gambling, drugs, unauthorised access into other computer systems, denial-of-service (DoS) attacks, data modification, embezzlement, identity theft, credit card theft, theft of trade secrets and intellectual property, extortion and terrorism. The latter has a slight overlap of activities which can be seen in Table 2.1. Both type of crimes consider exploitation of children in different forms: solicitation/prostitution and the second refers to content related crimes against children. Similarly, in 2006, two cybercrime classes emerged: TYPE I and TYPE II [93]. The former being mostly technological in nature; the latter, more pronounced human elements.

| author | year | classification |
|---|---|---|
| FBI<br>National Computer Crime Squad (NCCS)<br>Fraser [80] | 1996 | Intrusions of the PSN |
| | | Major computer network intrusions |
| | | Network integrity violations |
| | | Privacy violations |
| | | Industrial espionage |
| | | Pirated computer software |
| | | Others |
| UK Audit Commission<br>Audit Commission [16] | 1998 | Fraud for private gain |
| | | Theft of data |
| | | Unlicensed software |
| | | Private work |
| | | Misuse of personal data |
| | | Hacking |
| | | Sabotage |
| | | Pornographic material |
| | | Virus |
| Wall [251] | 2001 | Cyber-trespass |
| | | Cyber-deceptions and thefts |
| | | Cyber-pornography |
| | | Cyber-violence |
| Furnell [83] | 2001 | Computer-Assisted |
| | | Computer-Focused |
| James and Nordby [123] | 2002 | Computer as an instrument in criminal activity<br>Computer as a target of criminal activity |
| Gordon and Ford [93] | 2006 | TYPE I |
| | | TYPE II |
| Walden [249] | 2007 | Technology-based |
| | | Motivation-based |
| | | Outcome-based |
| | | Communication-based |
| | | Information-based crimes |
| Wall [250] | 2007 | Computer Integrity Crime |
| | | Computer Assisted Crime |
| | | Computer Content Crime |
| McGuire and Dowling [172] | 2013 | Cyber-dependant |
| | | Cyber-enabled |
| United Nations Office on Drugs<br>and Crime Vienna<br>[193] | 2013 | Acts against confidentiality, integrity and availability (CIA) of computer data |
| | | Computer-related acts |
| | | Computer content-related acts |
| Wall [252] | 2015 | Crime against machines |
| | | Crime using machines |
| | | Crimes in the machine |
| Tsakalidis and Vergidis [241] | 2019 | A: Offences Against the CIA |
| | | B: Computer-Related Offences |
| | | C: Content-Related Offences |
| | | D: Copyright and Related Rights |
| | | E: Combinational Offence |

Table 2.1: Cybercrime classification with the categories corresponding to CSEM highlighted.

Walden defined cybercrimes in five different categories: technology, motivation, outcome, communication and information based crimes [249]. McGuire and Dowling compressed the classification to two categories: dependent and enabled cybercrimes [172]. The former group defines a cyber-dependent crime as an offence that can only be committed with technology, e.g., live steaming of sexual abuse. The latter may be perpetrated without the use of ICT, but with its use, a large scale attack may be performed. For example, child sexual abuse material (CSAM) would fill in this category weather it refers to a magazine or a digital file.

The rest of the modern cybercrime classifications can be seen with the categories corresponding to CSEM related crimes highlighted in yellow. These classification are more sophisticated because they consider the principles of computer and internet security: CIA.

## 2.1.2 Digital Forensics

### 2.1.2.1 Introduction

In the early days, criminal investigations relied on oaths, confessions and witness testimony [226]. Verdicts were determined by inducing pain to the suspects, normally know as trial by ordeal. Later with the invention of computers, which eventually became very prevalent in society, being used for many task that would optimise manual processes. Nevertheless, there was also room for crimes enabled or dependent on computers. Born from computer hobbyists and law enforcement officers, the term computer forensics (CF) was coined. It was a term used in the past to what today we refer to as DF. Today, CF refers to the recovery and investigation of material only found on a computer. With the burst of technological advancements, evidence has expanded from personal computers to myriad digital devices and there has been an increase of sub-disciplines related to DF. According to the definition established in the first digital forensic research workshop (DFRWS) in the year 2001, DF refers to the use of proven methods for all the stages of the digital forensic process (preservation, collection, validation, identification, analysis, interpretation, documentation and presentation) applied to digital evidence [197].

### 2.1.2.2 Digital Forensic Process

Digital forensics processes are a set of standard procedures involved in a cybercrime investigation used to substantiate an event. They are an abstraction of practices used in real life investigations. Adapted from Kruse [144] the procedures involved in digital forensics are the preservation, identification, extraction, documentation and interpretation of digital data (PIEDI). First, preservation determines the act of maintaining the original evidence intact. Any change in the chain of custody would impact greatly and introduce doubt to a case. Second, identification of data with probative value is another challenge in an investigation. Police officers tend to collect as much evidence as possible and increase the volume of information to be analysed. Extraction is typically a copy of seized devices. Challenges in DF due to security mechanisms (that prevent forensically sound copies) in certain devices such as gadgets powered with android

OS have hindered digital investigations [12]. Following, proper documentation must be held from the beginning of the case. Any change in the state of the chain of custody must be documented. Finally, the interpretation of the data is processed by a software but interpreted by an expert.

According to Du et al., it is almost impossible to design a perfect model that can deal with any investigation [67]. For our research, we deal with cases where acquiring immediate clues are crucial. A digital forensic triage process model would seem adequate since they were proposed to deal with time sensitive cases [217]. Rogers et al. proposed a field approach for providing a swift response to the identification, analysis and interpretation of digital evidence in an efficient manner while optimising the acquisition phase. This model is named the Cyber Forensic Field Triage Process Model (CFFTPM) [217] and it is useful for cases regarding CSAM, missing and/or exploited children, where time is of the essence. The CFFTPM consists of 6 phases which can be seen depicted in Figure 2.1: planning, triage, user usage profiles, chronology timeline, internet and case specific.

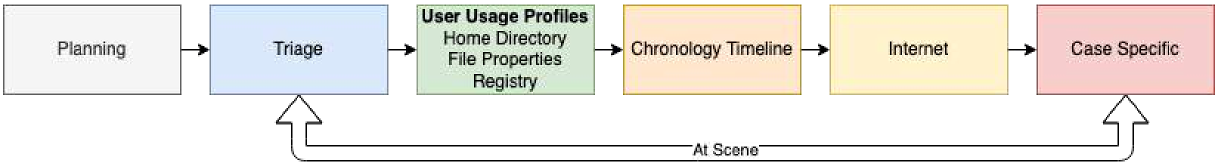


Figure 2.1: CFFTPM Phases

The first phase entails proper prior planning and the SALUTE[2] intelligence collection is recommended. The second phase refers to triage in which the elements are ranked in terms of priority. The 3rd phase mentions usage/user profiles. Although digital devices are more common to belong to a single user, there are other shared devices such as printers, home assistants and smart TVs that are communal between several users. The 4th phase recognises modification, access and creation time of a file. The 5th phase presents the internet related activities performed by the victim or suspect. The last phase refers to the specific case being investigated, the type of artefacts relevant to a case. Although there are cases that are a matter of life or death, the triage process phase should be handled with care. Legal trials that rely on digital evidence can be affected by superficial case analysis, lack of evidence integrity and validation, and not being able to present adequate reports [27].

[2] Strength, activity, location, uniform, time and equipment

#### 2.1.2.3 Process Models Applicable to CSEM

In a typical case of CSEM, the following occurs:

- Law enforcement finds CSEM on a digital device
- The visual media is sent to the prosecutor
- Charges are filed
- Suspect is arrested
- Defence attorney reviews the content
- Defendant takes a plea

In general, the first 48 hours of an investigation are critical to an investigation[192]. If a suspect is not arrested at a time, and CSEM is discovered after forensic evaluation of their digital media, then the suspect will be arrested later, usually in a time-frame of 6 to 12 months. This constitutes a danger due to the suspect being able to endanger other victims in the interim.

In 2002 there was a warrant-less seizure pertaining to the case of Charles Hinds, JR [49]. The officer McLean performed a consensual search of a computer for an email of interest. While searching on the defendant's home computer, CSEM was found. The officer seized the device and later-on obtained a search warrant. Thousands of CSAM were found and the defendant was convicted. It was found reasonable that the officer was able to seize the computer prior obtaining a warrant because the evidence could have been destroyed.

The procedure used by McLean can be seen depicted in Table 2.2. It can also be noted the detail and different phases of the digital forensic process employed. In this case it is evident that a Planning Identifying Preservation Acquisition Analysis and Presentation model was used, similar to the traditional process model presented by Rogers et al. [217].

It should be noted that the COMMONWEALTH vs. CHARLES HINDS, JR case emerged from a different unrelated case, from the family member John J. Hinds who was Charles Hinds' uncle and was investigated for murder. The officer was seeking for evidence that would support in the arrest of J Hinds but also ended up finding illegal material on a computer in the same network. One case of homicide ended up in two arrests.

| Action | Detail | Phase |
|---|---|---|
| Photo the system | Before dissasemble | Planning |
| Disconnect data and cables to the hard drives | Before moving equipment | Preservation |
| Place a write-protected boot disk in the disk drive | Write protection and boot control | Preservation |
| Label | What was seized and by whom. | Identification |
| Prevent electromagnetic exposure | While transported | Preservation |
| Logged/Bagged/Tagged | Every component | Preservation |
| New Technologies SafeBack | Image hard drive | Acquisition |
| Search existing and deleted files | DiskSearchII, TestSearch and IP Filter | Analysis |
| Present to court findings | 7 counts of possesion of CSEM | Presentation |

Table 2.2: Actions and the Respective Phase Taken for the COMMONWEALTH vs. CHARLES HINDS, JR Case.

In 2012, Boddington adopted a Toulmin's model of argument, which improved the efficiencies in the case management and communication with legal practitioners [27]. The Toulmin model considers six aspects for analysing arguments:

- **Data** which refers to the evidence (Subject possessed CSEM).
- **Warrant**, which is the component that links the data and the claim (Subject possessing CSEM is aware of it.).
- **Claim** which is the point of the argument, conclusion and sometimes called the thesis (Subject is guilty of possessing CSEM).
- **Backing** is the material that supports the warrant (proof of downloads, viewing, deliberate destruction of files, etc.)
- **Qualifier** is the soundness of the argument (Unless).
- **Reservation** is an exception to the claim. (Subject had an alibi for an action, Trojan horse defence, etc.)

Both approaches presented in this section give an overview on how the digital forensic process has been implemented in real cases. For the former, a traditional process model was applied; for the latter, the Toulin model for structured argumentation was used. It is worth mentioning that cybercrime and forms of abusing children has evolved through the years and in cyberspace. The models used today may not be applicable in future crimes. These models might not be consistent for live streaming cases of CSEM and the digital forensic process models should also adapt. This behaviour resembles the no-free-lunch theorem for ML and optimisation searches, where there is no one model that works best for every problem.

Finally, as stated by Rogers et al., cases of CAM should be given the highest priority [217]. Tools to assist visual media can aid the searching task; furthermore, directory structures, cookies, emails, temporary internet files and browsing history and downloads from peer-to-peer (P2P) networks should be considered in a smart manner. Therefore, the CFFTPM approach discussed in Section 2.1.2.2 should be applied and the triage phase should be implemented accordingly.

#### 2.1.2.4 Admissibility of CSEM Evidence

With the introduction and growth of modern information technology, the evolution of evidences held in court has migrated from just the 'traditional' physical evidence to the inclusion of digital evidence.

Digital evidence are extracted from electronic devices containing relevant probative information which are often seized from the suspect or given up as evidence by victims or witnesses.

The nature of the digital evidence must also go under scrutiny. Digital data are constantly being created, accessed and stored by multiple entities such as the person themselves through platforms within social media or through a third person such as closed-circuit television (CCTV) cameras. The form of which the digital evidence is presented can influence its reliability in court. An ATM bank transaction can be held in a higher account for a person's location than perhaps a person's social media location feed.

The admissibility of data with probative value requires confirmation that the digital evidence was obtained lawfully and is relevant to the case. Ensuring legal admissibility of digital evidence in a court is a challenging task. CSEM evidence is mainly composed of photo and video files. The visual media processing methods should be validated meticulously prior to them being handled to court due to the risk of the method not being accepted and further restricted for other cases [123]. Furthermore, any introduction of reasonable doubt may dismiss a case. To attempt to alleviate this issue, the Daubert standard is suggested. It was introduced in 1993 and has been used by most state courts in the USA as a rule of evidence to assess the reliability of scientific evidence through the following factors [184]:

1. The method can be and has been tested in the past,

2. It has been subject to peer review,

3. Error rates are acceptable, and

4. There is a general acceptance in the scientific community of the method.

In regards to these enumerated factors, Nutter states that "machine learning easily satisfies three of the four Daubert factors without extensive discussion" [191].

#### 2.1.2.5 Existing Tools and Techniques

French criminologist Dr. Alphonse Bertillon developed a system in the 1800s to identify criminals through drawings or photographs including the measurements of the face and body [101]. Later, his collection of criminals and facial features such as facial shapes, chin, hair, eyes, ears, nose, etc., was used to assist victims and witnesses who would provide descriptions to a forensic artist. This was that start of what would become the work of a forensic artist. Forensic artist techniques are relevant to kidnaps and missing children. In 2009, the National Center for Missing and Exploited Children (NCMEC) reported that of the nearly 26,300 runaways reported, 1 in each 6 cases were likely victims of child sex trafficking. Image modification has been employed by forensic artists. They have been asked to modify and age images to attempt to solve cases of missing children.

Digital forensic investigators require numerous tools and in certain cases, personalised scripts are necessary to automate any process involved in the investigation. Forensic soundness[3] is key in the use of digital forensic techniques. Forensically sound tools such as Encase and Autopsy [38] are widely used in LEAs throughout the world. Among the most rustic tools to preserve digital evidence are the hash comparisons and the bit-to-bit copy command. Both tools are forensically sound and commonly accepted in the digital forensic process. Currently there are a plethora of local and cloud-based tools to assist investigations. Nevertheless, the possibilities are infinite with the expansion of digital forensic sub-fields. For example, in network forensic investigations, Netcat (NC) "the TCP/IP Swiss army knife" could emerge as a first choice not only as a tool to assist network forensic investigations but also for performing live analysis. In any digital forensic investigation it is crucial to avoid writing to any mediums as possible. This becomes complex when live analysis is performed due to certain programs affecting not only the memory, registers but even the disk. NC enables the copying of files over a network while offering integrity. A tool that is capable of preserving captures of the network traffic in a live forensic scenario is elemental. Wireshark[4] is an open-source packet analyser used to monitor network traffic. It can also be used to detect conventional network attacks such as port scanning, covert application layer protocol such as file transfer protocol (FTP) or internet relay chat (IRC), Internet

[3] Forensic soundness is "The application of a transparent digital forensic process that preserves the original meaning of the data for production in a court of law [173]"

[4] `https://www.wireshark.org/`

Control Message Protocol (ICMP)-based attacks, DoS/DDoS, etc. [185].

Cloud-based tools such as CyberChef[5] "the cyber Swiss Army knife", are critical to perform "cyber operations" within a web browser. The web-based application created by the Government Communications Headquarters (GCHQ), allows the manipulation of data without using complex tools or algorithms. Several digital forensic-focused, linux-based OS have been released: Santoku[6] mainly for mobile forensics, Kali Linux[7], SIFT Workstation[8], Parrot Security[9], Caine[10], Pentoo[11], etc. Usually pre-built virtual machines (VMs) are available and ready to run without any installation required.

Finally, tools used to aid CSEM investigations are for example Autopsy with the corresponding ingest modules which are introduced in Section 2.2.6.1, forensic toolkit (FTK) is a court-cited digital investigations platform built for speed, stability and ease of use. The PhotoDNA tool is explained in Section 2.2.6.2 and is an automatic tool to detect CSEM. In 2010, d. Polastro and da Silva Eleuterio proposed the NuDetective Forensic Tool, which was developed to assist examiners. The aforementioned tool performs automatic nudity detection via image pixel analysis and file name inspection. The drawback of this approach is that there is not an implementation of an age detector and mainly relies on the filenames to judge the age, additional to a manual required revision. Furthermore, the filenames can easily be modified by the perpetrator and many illegal files would not be detected. Later in 2011, Ulges and Stahl introduced a CSAM system based on colour-enhanced visual word features and statistical classifications using support vector machines (SVMs) [242]. The system was able to assist binary classifications of illegal and non-illegal content by analysing small colour patches present in the image. Few researchers have proposed new tools and technologies to discover CSEM. One of them was Sae-Bae et al. who later in 2014 proposed the automatic detection of illegal content while improving on face and skin-tone detection [221].

#### 2.1.2.6 Digital Forensic Challenges

Connecting the dots accurately to solve a crime *against/with/in* a digital device, requires extensive time, knowledge, expertise and in some cases intuition. The role of intuition in research is to provide an "educated guess" which may prove later to be correct or incorrect. Moreover, if intuition is applied, further proof to backup the statement is

[5] `https://github.com/gchq/CyberChef`

[6] `https://santoku-linux.com/`

[7] `https://www.kali.org/`

[8] `https://digital-forensics.sans.org/community/downloads`

[9] `https://www.parrotsec.org/`

[10] `https://www.caine-live.net/`

[11] `https://www.pentoo.ch/`

required. With the vast amount of data that can be currently stored in digital devices, connecting multiple needles in a haystack is manually infeasible. The lack of experts and of automation to detect or prioritise evidence contributes in the accumulation of digital devices in digital forensic laboratories (DFLs).

Petabytes of information of our lives are consciously being transmitted daily from various social media platforms and communications via mobile messaging apps. Personal data is also being collected surreptitiously from location-based services, facial detection by private or state-owned CCTV cameras, and activities recorded by internet-connected sensors. All this data can be used as evidence and when collected for investigations are stored in DFLs awaiting to be processed. Practitioners have become overwhelmed by the amount of evidence they must deal with. A high percentage of investigations in LEAs are related with CSEM and the exposure to this content has been proven detrimental to practitioners[142].

Challenges for digital forensics are increasing with the appearance of new technology and numerous factors such as cloud environments that can not be analysed promptly due to the physical locations of the servers, complex deleted file recovery on solid state drive (SSD), volatility of memory, strong disk encryption and concealing mechanisms such as steganography that hamper evidence retrieval [123], pervasive Internet of Things (IoT) with lack of persistent storage; readily available tools that facilitate counter-forensics [181], omnipresent anonymous communications facilitated by Tor[12], big data, and lack of digital forensic datasets for research.

#### 2.1.2.7 Digital Forensic Backlog

Due to the proliferation of electronic devices, a person's life and personal information can now be found fully documented. A clear benefit can be observed with possessing large amount of data as it can drive correct convictions/persecutions at court; for instance, digital evidence obtained could refute a suspected person's alibi and instead locate the person to the crime scene. An intent of the offence may have been noted on social media sites which may denote the suspect's potential motivation for crime or perhaps even contain documentation of critical elements that proves the suspect guilty.

Conversely, whilst potential condemning digital evidence may be present or is believed to exist, obtaining such evidence may be impossible. There are laws such as The Fourth Amendment to the United States Constitution that protect the miss-use and unreasonable searches and seizures of such data [91]. Digital information belonging to the accused cannot be easily obtained without probable cause and thus preventing the ex-

[12] `https://www.torproject.org/`

traction of unrealised vital digital evidence. Often, such searches require the issuance of a warrant which may not be acquired in time of court proceedings. Furthermore, even when digital evidence is obtainable, the large dataset of information available to digital forensic investigators (DFIs) can be remarkably too significant as to create a data backlog that is not investigated/interpreted in time of the persecution. According to Quick and Choo, both increase of storage capacities in consumer devices and cloud storage services, and the number of devices seized by the law enforcement per case, has contributed to the accumulation of data [204]. Nevertheless, the significant lack of resources and automation in the acquisition and/or analysis steps of the digital forensic process increments the digital forensic backlog. Per Scanlon, the lack of these resources will continuously influence the throughput of DFLs and therefore, are likely to continue hindering DFIs in the future [225].

The digital forensic backlog is a common issue omnipresent in police agencies throughout the world. This problem not only affects the DFLs but has escalated to the courtrooms and although vast efforts to present processed evidence in a timely manner has been performed, the inevitable bottleneck preventing the release of information with probative value, has resulted in a significant delay in the judicial process. Henceforth, creating an impact on results in court cases being dismissed due to insufficient evidence [40].

Moreover, It has been almost a decade, since Casey et al. documented the issues of data backlog growth within DFLs. Delays from 6 months to 1 year had been recorded and ever since, the accumulation has increased exponentially. In 2016, as affirmed by Scanlon, the digital backlog can be delayed in an extreme of four years [225]. Today, we also have figures in the order of up to 4 years which were reported by multiple sources of Irish News, claiming that the Garda Inspectorate have reported delays going back four years.

Multiple approaches such as time-consuming / repetitive task automation , data reduction, digital triage and data de-duplication, have been done to attempt to alleviate the digital forensic backlog. However, the growing rates continue to increase.

#### 2.1.2.8 Triage

The term 'Triage' is believed to have originated as far back as the Napoleonic Wars through military surgeon Baron Dominique Larrey [215]. During the WWI, triage was reintroduced to the US military by the allied forces [126]. The triage process determines the priority of which patients are examined, based on the severity of their injuries. Those in need of more urgent and immediate care are primarily addressed [132].

In forensic investigations, handling digital evidence often requires the practice of triage to alleviate the evergrowing backlog. However, this practice must be handled with extra care in order to avoid loss of important data. With more digital evidence than there are efficient tools and workers to address them in time of court proceedings, triage is required to identify and analyse as much probative data as possible to aid the jury's final verdict.

In 2013, Overill et al. stated that the Digital and Electronics Forensic Service from the London Metropolitan Police Service, received more than 38,000 digital devices per year [196]. This value has long since been outdated and with the herald of new technological devices such as wearables and the IoT, its popularity and growth of storage capacity, the workload needed to examine all the evidence is enormous. Triage practices in DF would help sift through irrelevant data/devices and control the workload to a more manageable state.

## 2.2 Child Sexual Exploitation Material Investigation

### 2.2.1 Introduction

The number of internet users is constantly rising and each year increasing numbers of young people are online. The most vulnerable groups in cyberspace are subject to possible exposure to many variants of cybercrimes such as phishing attacks, hacking, sextortion, CSEM, child grooming, etc.

CSEM refers to the possession, creation and dissemination of indecent images depicting children. It is a digital content-related and information-based crime where child victims are explicitly portrayed engaging sexual activity. Indecent images of underage subjects are now prevalent due to the proliferation of data through the internet. CSEM increased with the internet while reducing the risk of detection of people hiding behind a chain of proxy servers such as mix networks, onion routers such as Tor and peer-to-peer networks such as Gnutella. "Cloud-computing technology similarly enables private access to storage that can host massive collections at a very low cost" [29].

The definition of underage (or child, minor, juvenile, etc.) may vary between jurisdictions, agencies/organisations, or even context for particular activities such as drinking alcohol, gambling, or qualifying as age of full legal responsibility. For instance, the United Nations Convention on the Rights of the Child [243] defines a child as anyone under the age of 18. For the purposes of this research, this definition of child/underage will be employed.

Child abuse investigations are common occurrences in LEAs throughout the world [8]. Possession of child abuse material has been in significant demand within investigations [130]. These investigations have become an arduous task due to the increasing usage of anonymization tools, private P2P networks [115], and cloud-based KVM systems [75]. Worldwide law enforcement and child protection communities have been struggling to diminish CAM and combat human trafficking. Organised criminal groups are operating in the deep web, which is a hub for criminal black markets, where paedophiles are able to exchange vast number of CSEM; often to obtain acceptance within a group of paedophiles and ultimately gain access to other collections of illegal content [143]. Although there is a huge computer crime industry impacting billions worth to the economy, the effects caused on victims of CSEM is matchless. Abuse and sexual exploitation have lifelong repercussions on child victims. Both physical and behavioural problems are developed.

### 2.2.2 Prevalence

In 2019, the NCMEC's CyberTipline received 16.9 million reports related to suspected CSEM. From all these reports, a total of 69.1 million visual media files were extracted. The major platform reported was Facebook representing a 94% of the 16.9M. It is assumed that the reports are included from Instagram, Whatsapp, etc., due to PhotoDNA being deployed on their entire network [74]. However, it is unclear if PhotoDNA works for Whatsapp due to the current end-to-end encryption technology. The next biggest submitter was Google with a total of 449,283 reports and Microsoft (deployed in both OneDrive and Bing search engine) with an amount that equated to 123,839. The top three companies (Facebook, Google & Microsoft) used the PhotoDNA which is further discussed in Section 2.2.6.2; the utility was donated to NCMEC, has been implemented in over 155 organisations and has been responsible for removing over 10 million CSEM images without dispute [74].

The presence of CSEM can be related to the amount of reports received per year. Nevertheless, not all content gets reported for various reasons such as the victim not disclosing the assault, the content being concealed within small groups for a long time, and the difficulty of handling and regulating material when it cannot easily be determined that minors are involved. The NCMEC's CyberTipline documented reports by electronic service providers (ESP) for 2009. The values can be seen depicted in Figure 2.2. This figure excludes the top three companies previously mentioned. The image sharing services had higher number of submissions than other services. These include Snapchat (82,030), Imgur, LLC (73,929), Twitter, Inc. (45,726) and Discord Inc. (19,480).

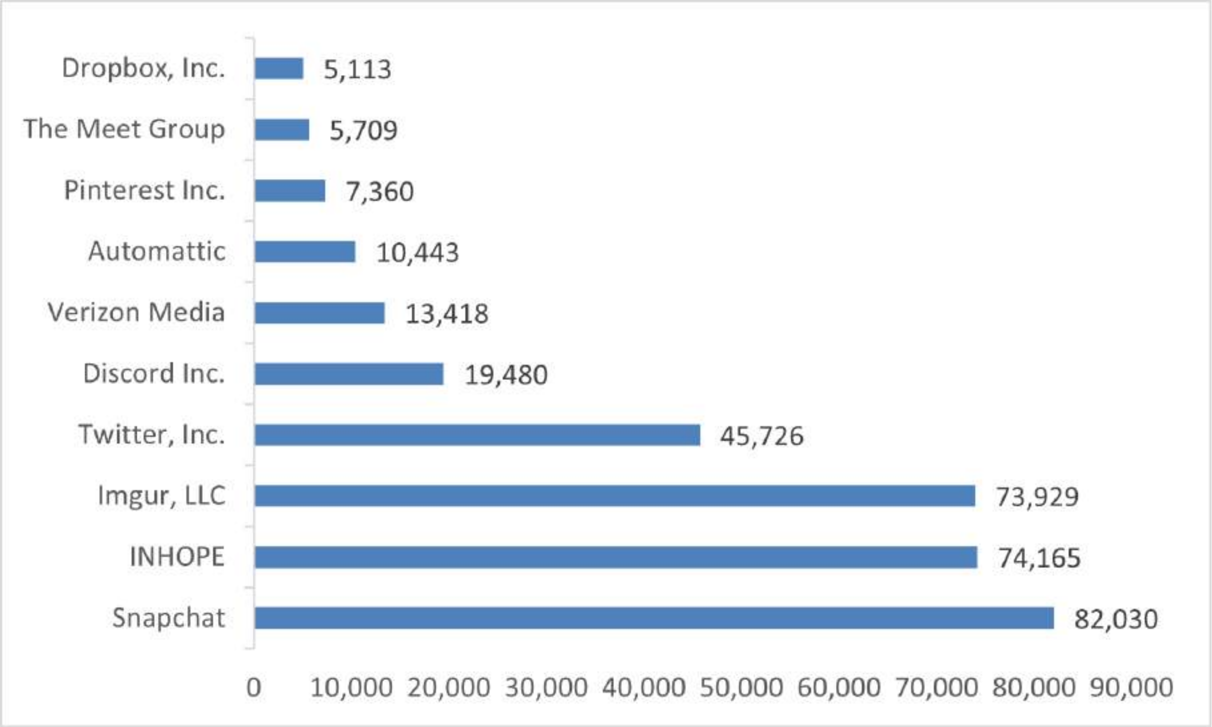


Figure 2.2: Reports by ESP without considering Facebook, Google & Microsoft

The amount of reports per year constitute an alarming statistic that should raise awareness of the public and take the cases of CAM seriously througout the world.

### 2.2.3 Cost

Criminals choose to remain anonymous in cyberspace and avoid using credit card payments to risk identification. Therefore, cryptocurrencies such as Bitcoins and CAM itself has become a currency, being often used as a means of payment to access other illicit material [194]. Back in the mid 1970s, a magazine containing CSEM would have cost approximately $10. Today, the whole content would be available for free [127]. Furthermore, the majority of EU law enforcement specialists have predicted that there are still a minimal amount of people who still pay for CSEM [73]. Nevertheless, revenue of $1.4 million in a one month span is the largest known commercial CSEM case [22]. It has been reported by the EUROPOL that a single video file on demand of new material content could cost as much as $1,200 [73].

### 2.2.4 Terminology Used in the Literature

Illegal material depicting minors has used several terms in the literature ranging from child pornography, crimes against children, sexually exploitative imagery of children

(SEIC) [4], indecent images of children (IIOC) [183], CAM, CSEM, CSAM etc. The use of dissimilar terminology makes the comparisons across studies strenuous [116].

There is a direct requirement for LEAs and researchers in general, to adapt the same lingual terminologies, specifically in the context of suspect and victim identification. Normalised data allows cooperation and integration through datasets collections both nationwide and globally.

In 2016, the Luxembourg guidelines were created, also know as the terminology guidelines for the protection of Children from sexual exploitation and sexual abuse [238]. These guidelines established normalisation for the complex lexicon of terms that are commonly used when addressing the sexual abuse and exploitation of children and are highly recommended for LEAs across the globe. The terms "child sexual abuse material" and "child sexual exploitation material" should replace the term "child pornography (CP)". Pornography is a term used for adults performing consensual sexual acts. This term does not apply to children because there is no consent. Nevertheless the term CP is still used predominantly not only in academic papers but also when addressing legal issues and contexts [238]. This term should be avoided in a non-legal context.

The European Parliament has corroborated the need to use the correct terminology for the aforementioned crime and use the appropriate term "child sexual abuse material". Although the term CP is consistent with the terminology used in Article 2 of the Optional Protocol to the Convention on the Rights of the Child on the Sale of Children, Child Prostitution and Child Pornography, 2000, the terminology used in our study is adhered to the Luxembourg Guidelines.

### 2.2.5 Modus-Operandi

Cybercrime is opportunistic in nature. Overill and Silomon discussed that cybercrime may be viewed as an unbalanced conflict – where the resources required to mount a decent defence are much greater than those required to execute a successful attack [195]. This resembles terrorism and guerrilla warfare. For cyber-dependant crimes involving minors, the asymmetrical behaviour is similar. The resources needed to produce illicit content are technological devices that can create, store and distribute material. However, the distribution is not that simple. Perpetrators who wish to cover their tracks are involved in closed groups with contacts that are well known throughout their community. Jenkins et al. documents that some individuals would have contributed over a period of 5 years and would be reluctant to trade with any other unknown members. The resources needed to find crimes involving the possession of CSEM may resemble seeking a needle in a haystack. Often requiring extensive days of research on the internet – potentially on newsgroups, anonymous image-board websites, IRCs, onion networks, Freenet[13], etc.

As mentioned in Section 2.2.3, CSEM was found widely available in magazine stores in USA back in the mid 1970s. Recently, in 2018, it was documented that there is still content suspected of being CSEM in physical retail stores in Tokyo's Akihabara district. The Human Rights Now (HRN) is a Tokyo-based international human rights non-governmental organisation (NGO) that has confirmed the existence of CSAM sold in Japan [190]. The products in retail stores have not passed any relevant reviews and lack of a system that confirms the age of the performer depicted in the pornography scene. There are three main issues of the aforementioned material in the stores: 1) many of these products are advertised openly as if they contained CSEM, 2) some of the storage devices such as CDs and DVDs that contain pornographic material advertise that their actors are under the age of majority, and 3) some do not advertise underage actors but the age of the performers are physically apparently minors.

In cyberspace, the hosting of CSEM resembles guerrilla troops – it cannot have an established or permanent base and it is set in a stealthy manner attempting to avoid the authorities. The material is usually accessible in private servers with hidden services that use Tor technology to stay secure and anonymous – only accessible to tech-savvy individuals who also share the same kind of content and are members of the closed group. That said, it is a network that is hard to penetrate even by law enforcement. This is quite a severe technique because it hampers criminal investigations and forces authorities to break the law if they wish to pursue the investigation. It should not be forgotten that each time illegal content is accessed or shared, a victim is re-victimised. Even the most innocent nude images could be subject to molestation – once out in the deep waters of the dark-net, there is little that can be done to eradicate such unlawful material.

On one of the posts documented by Jenkins et al., an anonymous commenter states that novice users are mainly the ones that have been charged with CSEM-related crimes while sending their broken computer to a repair shop and being reported by the technician, or contacting an undercover officer while exchanging material via email and/or IRC [127]. Nevertheless, the grand majority of cases taken to court are product of other unrelated offences such as molestation.

[13] Freenet is a P2P platform for censorship-resistant communication. It uses a decentralised distributed data store to keep and deliver information.

### 2.2.6 Approaches to Identify and Combat CSEM

Due to the ease of which CSEM can be distributed it becomes impossible to completely eradicate an illegal file once it has been shared. LEAs devote their valuable time seizing harmful websites, material and attempting to prevent the redistribution of known CSAM. Nevertheless, file duplication, file tampering, steganography, backups in private cloud services and new content generation is almost impossible to be anticipated.

International law has increasingly recognised children's rights and groups have been formed to assist in the combat of ICT-facilitated child sexual abuse and exploitation [194]. Unfortunately, there are still around 53 countries that still have no law at all that specifically criminalises CAM pornography [5]. On the technological side, several approaches have been considered to identify unlawful material. An early approach was the manual inspection of storage devices but with the effect of Moore's law, this approach is not feasible anymore without an automated tool. Cryptographic hash functions were proposed and with the emerge of AI, several techniques have been suggested.

#### 2.2.6.1 Traditional Hash Database Approach

A hash is a one way function that converts information of any given size to a fixed sized structure with the use of secure hash algorithms (MD5, SHA-1, SHA-256, SHA-512). The first tool of choice when investigating sheer volumes of data is the use of hashing [219]. Almost every examination of seized evidence is subject to pervasive hash searches [222]. Digital forensic examiners use hash functions throughout the forensics process, from the acquiring and analysis phase to the presentation of evidence in court. Hashes can be used to preserve evidence while keeping a record of the original hash, versus the bit-to-bit copy. These hashes should match, otherwise there is a compromised integrity in the chain of custody which could be questioned in court and would contribute to the case being delayed or dismissed. Another application of hash functions is to filter good-known files or bad-known files. While performing a search on storage devices such as a HDD or SSD, to perform the task in a shorter time, these filters would be considered. There is no point on scanning files that belong to the operating system or office suite files for example. Certain ingest modules for Autopsy[14] support triage-related features such as prioritisation, file filters and ingest profiles. Autopsy may create ingest profiles that search only pictures; use the hash lookup module which

[14] Autopsy is a digital forensics platform and graphical interface to The Sleuth Kit and other digital forensics tools.

calculates MD5 hash values and compares them to a database such as NIST National Software Reference Library (NSRL)[15], or employs known child exploitation hash sets from LEAs.

#### 2.2.6.2 Improved Hash Search Approach

The disadvantage of using fixed child exploitation hash sets as specified in Section 2.2.6.1, is that any minor change on a photograph, geometric or photometric transformations such as rotating, scaling or noise addition would invalidate the matches that the ingestion module would produce. Moreover, most online services that support image upload, automatically modify the original images. In 2018, Farid mentioned that Facebook and Twitter platforms would usually resize, recompress and strip certain metadata from the uploaded images [74]. In 2009, PhotoDNA was launched to address simple modifications such as re-compression, resizing, colour changes, and labelled text. PhotoDNA is a Microsoft-based technology used to identify similar visual media (particularly CSEM) through the computing of hash values. It uses a perceptual fuzzy hash technology [183].

An overview of the algorithm implementation of PhotoDNA is discussed in [74] and we have depicted the steps in Algorithm 1 to compare with our hashing approach which can be seen in Algorithm 3 and explained in Section 5.5.4.7.

**Algorithm 1** Robust-Hashing - PhotoDNA

```
procedure HASH_COMPARE($image$)                    ▷ Returns a boolean
    Convert $image$ to grey-scale
    Resize $image$ to 400 x 400 pixels
    Apply high pass filter to $image$
    Partition $image$ into $quadrants$
    Extract feature vector $p$ from $quadrant$
    Compute euclidean distance between feature vectors (Equation 2.1):
        $$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2.1)$$
    while $true$ do                    ▷ Iterate until end of feature database
        if $d(p,q) < threshold$ then
            Return $true$
```

First, a full size resolution colour image is converted to grey-scale and resized to a fixed resolution of 400 x 400 pixels. Next, a high pass filter is applied to highlight

[15] The NSRL is designed to collect software from several sources and include computed file profiles the software into a reference dataset.

the salient image features. Then, the image is partitioned into quadrants for which statistical measurements are applied to form the hash. Finally, the euclidean distance as shown in Equation 2.1 between two feature vectors $p$ and $q$ in an Euclidean *n-space* is calculated. A threshold is defined and if the calculation falls below the specified threshold, there is a match. The loop depicted in line 8 produces a linear complexity $O(n)$ and could potentially be improved to a logarithmic complexity.

Another approach that addresses limitations of traditional blacklist-based approaches is approximate matching, which is able to deal with cases of merged, embedded, partial and modified files. In 2018, Lillis et al. proposed a hierarchical bloom filter tree (HBFT) data structure approach focusing on the MRSH-v2 algorithm that compresses any byte sequence and outputs a similarity digest. This approach lessened the running time of matching between collections while maintaining effectiveness [157].

#### 2.2.6.3 Machine Learning Approach

Identification of CSEM with the use of ML can be addressed either as a single problem or an ensemble of models. Due to sensitive and unlawful content, the former approach is unfeasible for this study. The latter approach has been taken into consideration. Usually automated CSEM detection systems consist in two components: a nudity detector such as the techniques developed by Sae-Bae et al. [221] and an age estimation model.

A pipeline that ultimately detects CSEM can be seen depicted in Figure 2.3. The nudity detector is initially used to flag and filter images and be further used in conjunction with an ensemble of models that endeavour to solve a bigger problem which is the detection of illicit material related to minors. The nudity detector can be as simple as a binary classification problem that outputs if the image contains nudity or not such as the ones developed by [15],[51], but it can also be more sophisticated by aggregating classes and classifying types of nudity, such as the work published by Sevimli et al. [229] in which there are five different classes: normal (class 1), swimming suit (class 2), topless (class 3), nude (class 4) and sexual activity (class 5).

Followed by the nudity detector is the underage age detector. The age of a subject can be judged by the face, signs of puberty, physical development, skin firmness, jawline, foreheads, cheekbones, neck, facial hair, etc. Each one of these components could constitute part of an ensemble of models to predict age by averaging the results. Nevertheless, not all of these components would be available at a given time. In this dissertation, we have focused on the underage facial age detector module with a multi-class (classification) or single output (regression). A binary classification class is avoided because if the output is limited to a true or false, the whole training/validation/testing proced-

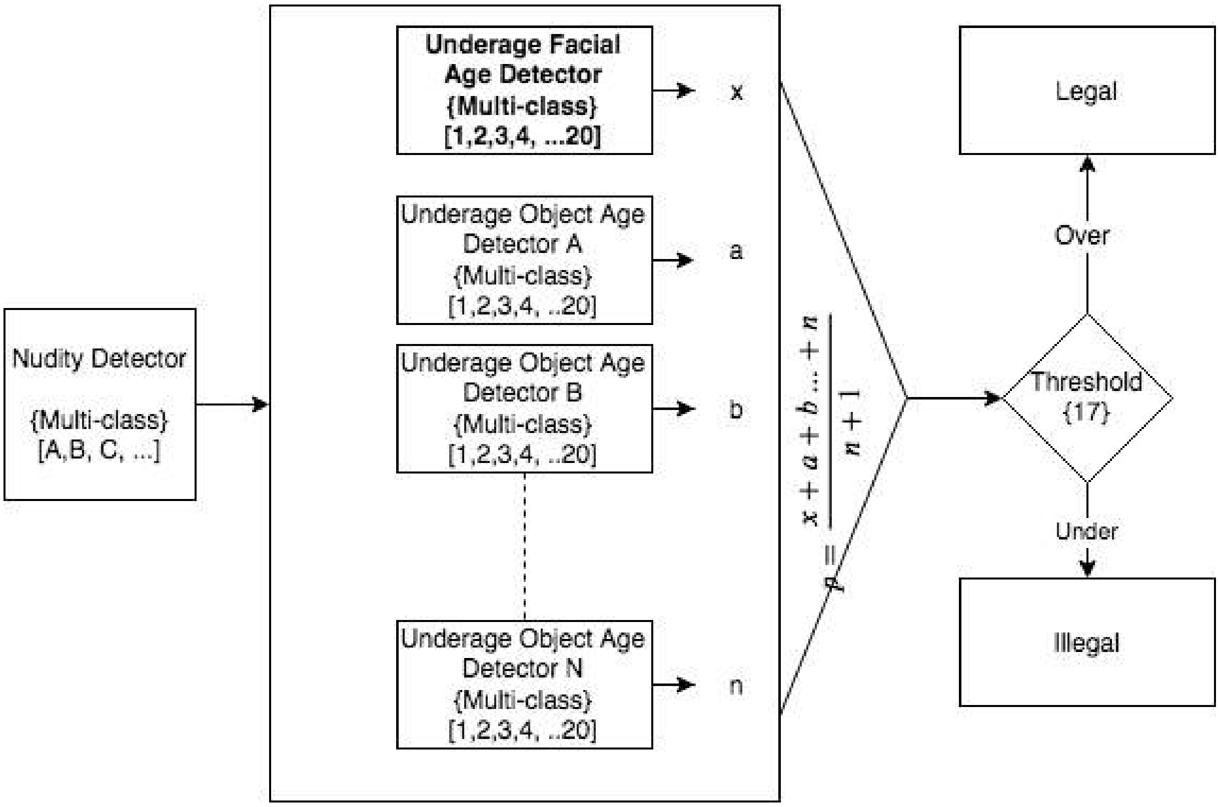

Nudity Detector
{Multi-class}
[A,B, C, ...]
Underage Facial Age Detector {Multi-class} [1,2,3,4, ...20]
x
Underage Object Age Detector A {Multi-class} [1,2,3,4, ..20]
a
Underage Object Age Detector B {Multi-class} [1,2,3,4, ..20]
b
Underage Object Age Detector N {Multi-class} [1,2,3,4, ..20]
n
$p = \frac{x+a+b...+n}{n+1}$
Threshold {17}
Over
Under
Legal
Illegal

Figure 2.3: Pipeline of Designed Ensembles to Detect CSEM [10]

ures would be fixed to a hard-coded threshold separating underage from adulthood. This would be inconvenient and incompatible for certain countries where the upper bound age of the definition of a child is different than eighteen. In the case of the Republic of Ireland, the definition of child is anyone under the age of 18. In 1999, New Jersey's child pornography statute defined a child as "any person under 16 years of age" [259]. Finally, the state-of-the-art MAE (as explained in Section 2.4.4) for age estimation should be considered to establish the threshold. For instance, if an age lies in the borderline between child and adulthood, a threshold of 15 years should be considered for children due to failure of 3 years being plausible. Once the threshold is set, it is a matter of mapping the output to a result of either *legal* or *illegal* content.

### 2.2.7 Identification of Subjects via Visual Media

In 2016, the NCMEC received 8,000,000 reports of CSEM, 460,000 reports of missing children, and 220,000 reports of sexual exploitation [74]. These worrisome figures that LEAs and organisations have to deal with constantly, show that CSEM has been reported the most. Moreover, missing teenagers are at high risk of becoming victims of sex trafficking. In the Republic of Ireland, the Garda Missing Persons Bureau maintains accurate records on missing persons within the jurisdiction, assists investigations of missing persons, aids the identification of bodies and administers photographs of missing persons on the following websites: `www.garda.ie` and `www.missingkids.ie` [177]. For cases of missing children outside of the jurisdiction Interpol is assigned. Furthermore, Schengen Information System (SIS)[16] in efforts to prevent and assist content-related crimes, have created alert categories. The purpose of the alert category is to find missing persons, particularly children, and place them under protection if lawful and necessary.

It is import to be able to identify a subject through images or videos. Whether to identify an offender/suspect or a victim of ransom, child solicitation or exploitation. Individuals have been identified without their consent in the past with visual media such as images and videos from private CCTV cameras, state owned surveillance devices, etc. It was reported in 2019 that U.S LEAs were using motor vehicle records to identify American citizens without their consent; similarly, in China, the state police has used facial recognition technology to police behaviour [232]. Back in the early 2000s, the police started to use recognition technology but the accuracy was low, leading to unreliable systems. Due to AI technologies, significant advances in facial recognition have followed [1]. LEAs in several states such as Illinois, Texas and West Virginia, have acquired sophisticated camera surveillance systems that are capable of capturing and identifying faces in real-time [86]. Furthermore, studies estimated that 117 million adults are already in face recognition datasets used by law enforcement [162]. Although these datasets are used to assist criminal investigations, there are ethical and privacy concerns.

## 2.3 Dataset Curation

Dataset Curation for facial age estimation is the performance of selecting, organising, and storing facial images in a specific data collection. It involves the labelling and dissemination of data such that the value of the data is conserved over time, and the data continues available for reuse. Images should be curated and are predominately frontal face photographs of a single subject. Automatic facial and attribute detection can assist this process in a swift manner. Exposure, occlusion, noise, emotion and skin tone are influencing factors on the accuracy of underage facial age estimation [9]. As a result, images should be discarded according to the level of some of these factors, thus decreasing the problem to a smaller one. But care should be taken to include a more

[16] The SIS is the most widely used and largest information sharing system for security and border management in Europe.

diverse set of images.

### 2.3.1 Data Selection

Selection of a facial image is driven by the quality and the type of problem that needs to be solved. The main focus of this research is to address the digital forensic backlog by evaluating the use of tools that could automatically analyse underage photos. The collection, possession, and dissemination of CSEM is illegal. Therefore, the underage component is selected for studies whereas the nudity component is completely isolated and not considered. The selection is a group of underage single face images. However, certain dataset sources contain low quality images, photos with noise or multiple faces. The selection of high quality single frontal images with creative commons (CC) licenses is highly recommended.

### 2.3.2 Data Collection

Several online sources are available for facial age estimation. The predominant age labelled datasets discussed in Section 3.3 are a good start and other sources of images with age labels are available. Social Media is not recommended due to copyright concerns but there are web sites for online photo management and sharing such as Flickr[17]. Web crawling is a technique that has been widely used for research. It is an automated collection of data from web pages. The first crawler was coded in 1993 and is almost as old as the World Wide Web [106].

### 2.3.3 Data Organisation

The structure of the dataset is set according to the problem it tackled, and also depends on the size of the dataset. It is simpler to have all the files in a single folder and a metadata file containing the labels of the images mapped to the filename. In a single-folder structure, it is possible to iterate the folder with a script and further classify the images in separate folders for training/testing/validation. An example of binary classification organisation can be seen depicted in Figure 2.4

In a regression problem, it is better to maintain the files in a single folder, the filenames should have a specific nomenclature that would allow easy parsing and finally apply stratified shuffle split whenever needed. The shuffling technique applies stratified ran-

[17] `https://www.flickr.com/`

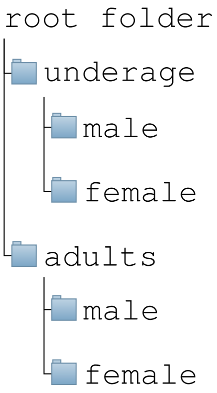


Figure 2.4: Binary Classification Structure

domised folds. These folds are made by preserving the percentage of samples for each class. Refer to Figure 2.5.

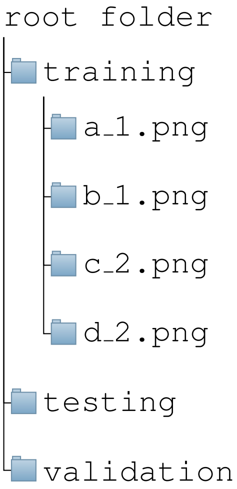


Figure 2.5: Regression Structure

### 2.3.4 Dataset Evaluation and Validation

Dataset evaluation can be performed on specific tasks such as classification, regression, detection, etc. (outlined in Section 2.4.2). It can provide transparency and accountability over facial recognition and age estimation related systems. The evaluation is an integral part of the dataset development process and it helps to find the best suitable model. Moreover, standardised dataset evaluation is paramount to allow the comparison of results from several other models. The main difference between dataset evaluation and validation is that the former compares the dataset with other datasets and the performance they produce with a given model. Whereas the latter evaluates the quality of the data inputs.

The dataset may be validated manually or automatically. For the former procedure, each image is checked for facial partial completeness (two eyes, a nose and a mouth). This procedure is usually applied to small datasets (Refer to Section 3.3.2). Therefore, small datasets can be employed mainly for testing and validation. The main difference between validation and testing is that the validation dataset is used during training to select the best classifier. Whereas the testing dataset is used to benchmark the final model that was fit on the training dataset. For the latter procedure, facial image averaging is useful to have an insight of what data was processed and if there is any anomaly within the dataset collected. Age and gender estimation models can be used to filter data. But due to performance issues, the use of data age range groups and a given threshold is preferred. Automatic gender detection has achieved higher accuracy in the past, but certain age ranges such as the ones close to the lower age range are still challenging. These automatic methods can alleviate the huge load that human labelers experience. Another dataset validation technique can be the use of crowd-sourcing to label information such as age, gender and ethnicity.

### 2.3.5 Data Storage

Recently, large data containers are advertised in cloud-based platforms, Amazon S3 is an object storage service composed of computer systems distributed across multiple data centres around the globe. Google Drive, One Drive and Dropbox are cloud based storage applications that have been used to share information. Large datasets have found place in these storage instances and now are commonplace in cyberspace. Metadata can be stored in spreadsheets, plain text, binary data container formats for MATLAB users, sqlite files, relational and non-relational collections, etc. Storing images in a database table is not recommended due to the size of the information and time to process huge amounts of traffic. An alternative, and better method is to store the images outside of the database and store only a link to the image file.

## 2.4 Machine Learning

### 2.4.1 Introduction

The trends of AI/ML/DL have been increasing rapidly over the past 5 years. This growth can be seen depicted by the Google Trends comparison in Figure 2.6, where the predominant tendency is observed on the ML line followed by AI and DL.

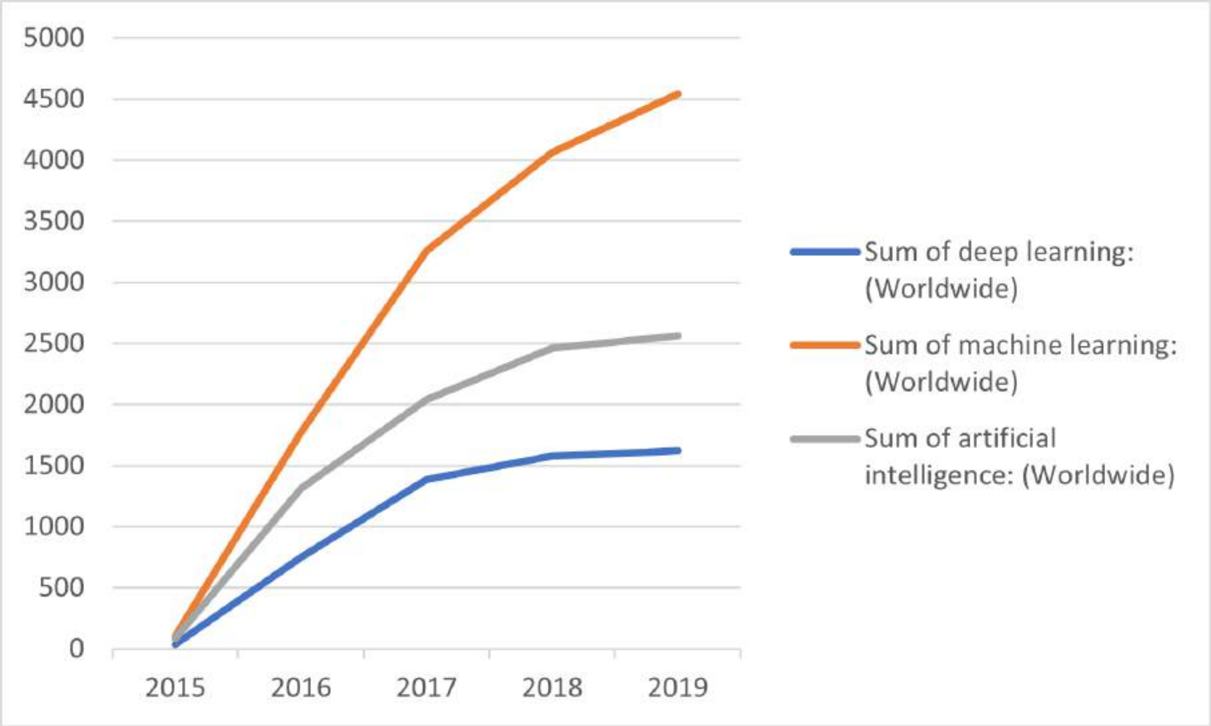


Figure 2.6: Interest over Time of Machine Learning/Artificial Intelligence/Deep Learning (The numbers represent the sum of the search interest per term, i.e., a value of 100 is the peak popularity for the term, 50 meaning it is half popular and 0 not popular) [92].

The inclusion of DL as tools and techniques to detect, gather and process content-based evidence in a timely manner is proposed. The age estimation problem mentioned in Section 2.2.6.3 is the centre of the study that can be addressed with DL which is a specific kind of ML that is able to categorise the aforementioned problem in several tasks such as classification, regression and anomaly detection. The tasks are described in Section 2.4.2.1, 2.4.2.2 and 2.4.2.3 respectively. Supervised learning (the main focus of this research) which "learns a function that maps an input to an output based on example input-output pairs" [220]. For cases of CSEM and missing children, missed detection can be harmful. It is important to take into account the speed and accuracy of these methods and how to improve them.

ML basic concepts such as generalisation, curriculum learning (CL), knowledge distillation, transfer learning, fine tuning, ensemble learning and finally generative adversarial network (GAN) are introduced. These subtopics are used throughout the research.

### 2.4.2 Machine Learning Tasks

As stated in [90], ML allows us to tackle tasks that are infeasible or too complex to solve without a human programmer. ML tasks are how the ML system should process the output of a feature. ML tasks such as classification and regression are categorised as supervised learning algorithms because the dataset contains items that are fully labelled. Whereas anomaly detection is an unsupervised learning algorithm that learns useful properties of the structure of the dataset while not requiring the items to be labelled.

#### 2.4.2.1 Classification

Classification is a supervised machine learning task. Given a set of input features, a classification task specifically refers to the prediction of an output that represents a class. For example, in an age prediction task the inputs can be an image of a face, where the pixel values in the image are the feature inputs and the output would be a specific class such as child, teenager, or adult. According to Goodfellow et al., modern object recognition is another example of classification. It is best completed with DL and implements the same basic technology that enables models to recognise faces [90]. Goodfellow et al. defines the classification output as a function $f : \mathbb{R}^n \rightarrow \{1, .., k\}$, where the function $f$ with domain $\mathbb{R}^n$ is the set of real numbers, with $n$ input variables and a range 1 to $k$, where $k$ is the number of classes.

#### 2.4.2.2 Regression

Regression is a supervised machine learning task. Given a set of input features, a regression task specifically refers to the prediction of an output that represents a continuous value. For example, in an age prediction task the inputs can be vectors representing facial features and the output would be a continuous value representing the age as a float. It may be argued that age fits into the classification task category. Nevertheless, as the age being able to be divided into months, a more specific value can be presented. Goodfellow et al. defines the regression output as a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ [90], where the function $f$ with domain $\mathbb{R}^n$ is the set of real numbers, with $n$ input variables and a range $\mathbb{R}$ which represents the output (a real number).

#### 2.4.2.3 Anomaly Detection

Anomaly detection is an unsupervised machine learning task. Given a set of objects or events, an anomaly detection task refers to the prediction of atypical behaviour; it is often applied on unlabelled data which is commonly known as unsupervised learning. A typical example is the anomaly detection in network traffic. Anomaly detection may be employed in web layer evolutionary-based packets classifications [141] that may assist CSEM investigations. Implementing underage detection as an anomaly detection task may require further research. While the problem would search for underage images in a large set of images, there are subjects that are very hard to differentiate between child and adulthood. Specifically when they lie in the borderline between both classes. Thus, not differing significantly from the majority of the data. Additionally this approach is mainly used when there are few samples in certain class which constitutes the anomalies. In age estimation, there could be millions of samples for either classes, for which a supervised learning task approach would be advised.

### 2.4.3 Generalisation

Generalisation as per the Oxford dictionary, is a general statement that is based on only a few examples. Humans tend to generalise with their own experience, experience of close contacts, samples from media, etc. In machine learning it is the ability of a model to adapt adequately to new, previously unseen data. It can be observed when testing a model not on the training data but on the data that it has never seen before. The model learns to make predictions based on training data. The more training data, the less chance of overfitting. As of 2016, Goodfellow et al. would advise that a supervised DL algorithm would achieve acceptable performance with 5000 examples per class [90]. Fernández et al. stresses in their work of regression-based facial age estimation, the importance to develop methods that can exploit large databases to gain substantial generalisation capabilities [78].

With training data, the outcome is already known and hence a 100% accuracy could be achieved. Nevertheless it is meaningless to do such comparison. The predictions from the model and known target values are compared, while the model's parameters are changed until both line up. The main reason of training is to develop the model's ability to generalise in a successful manner. The ability to generalise well for universal age estimation models is challenging. Factors such as environment, habits, ethnicity, makeup, etc., affect directly the age of a person. It is recommended to tackle a smaller problem by imposing limitations to the age estimation model so it can generalise better for the controlled dataset or use a CL approach (Section 2.4.5.1).

### 2.4.4 Evaluation Model Metrics

Evaluation model metrics explain the performance of a model. The metric depends on the task and can be measured by accuracy or error rates. Classification tasks are measured by accuracy, recall, precision, F1-Score, etc. Conversely, regression tasks are measured by mean squared error (MSE)/root mean squared error (RMSE), MAE, $R^2$, Adjusted $R^2$, etc. Accuracy is calculated by the proportion of samples that produce a correct output. Whereas the error rate is calculated by the proportion of samples that produce an incorrect output. The generalisation loss is usually the value we are interested in because it tells us how well the model performed with the data it has not seen before.

The top 3 evaluation model metrics used in our research are accuracy, MSE and MAE. The last two are error rates widely used as regression losses in ML. The former is the most commonly used loss function for regression. The latter is a more human-friendly measurement used to measure the average of the absolute mean error between the ground truth and the predicted values.

Finally, mean absolute difference (MAD) is the average absolute difference of two random variables $X$ and $Y$ independently and identically distributed. The formula is shown in Equation 2.2. This measure of statistical dispersion was used to calculate the performance per age.

$$MAD = E[|X - Y|] \tag{2.2}$$

Equation 2.3 depicts the accuracy for machine learning. Where $TP$ refers to True Positives, $TN$ is True Negatives, $FP$ refers to False Positives and $FN$ is False Negatives. This equation is further used in Section 6.2.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Equation 2.4 depicts the loss regression function MSE. Where $x$ refers to the ground truth, $y$ is the predicted value, and $n$ is the length of the set of values.

$$MSE = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{2.4}$$

Equation 2.5 depicts the MAE; where $x$ refers to the ground truth, $y$ is the predicted value, and $n$ is the length of the set of values.

$$MAE = \sum_{i=1}^{n} |x_i - y_i| \qquad (2.5)$$

### 2.4.5 Optimisation Strategies

#### 2.4.5.1 Curriculum Learning

Many times researchers have been encountered attempting to improve model performance by applying countless approaches without thinking to improve a simpler problem. The idea of CL or shaping is to start small and learn the basic aspects of a task to gradually increase the difficulty. CL was proposed by Elman in 1993 and was shown to improve network performance for several tasks and is used to speed the convergence while achieving better generalisation [23]. Per Goodfellow et al., it can be interpreted as a continuation method [90]. CL has been successful on a wide range of computer vision tasks [145, 150, 235]. In 2020, Buyuktas et al. proposed a CL approach for face recognition. The training set was subdivided into sets that would increase the difficulty based on the head pose, yaw, angle pitch and roll angles (These alterations have been proven to be influencing factors on age estimation in our work described in Section 3.2.11). The accuracy of this approach had better performance than using random batches [36]. The same approach can be considered to improve facial age estimation by creating a subset with only frontal faces and then introducing difficulty gradually.

#### 2.4.5.2 Knowledge Distillation

Model compression or knowledge distillation (KD) is the process of transferring knowledge from a large model (teacher) to a less computationally expensive smaller model (student). Usually the teacher's outputs (last or a hidden layer) are used as the student input's, unless data augmentation techniques were employed. KD has been proven to be very effective not only in training a student model but also in improving and sometimes outperforming a teacher model [198].

Research on KD is ample. A recent age-estimation-related approach used KD to create two teacher models. The basic idea was to transfer the ordinal knowledge captured by the first teacher (ranking model) and the dark knowledge captured by the second teacher (the multi-class classification model) to a compressed and less computationally expensive student model [271]. This research resulted in a more accurate performance in comparison with other state-of-the-art methods.

#### 2.4.5.3 Supervised Pre-training and Transfer Learning

Pre-trained models are models created to solve a specific problem but are reusable for similar problems, and according to Goodfellow et al., may help both in terms of optimisation and generalisation [90]. A supervised pre-trained model is a model that was trained on a large dataset with labels and has a similar problem to be solved. Per Goodfellow et al., strategies that involve training simple models on simple tasks before training the desired model to perform the desired task is known as pre-training. Due to time and expenses, it is common to use pre-trained models from the literature. For example, ImageNet is a large-scale image database built upon the backbone of the WordNet structure with an equivalent of tens of millions of annotated images [60]. This dataset has been used by well known convolutional neural network (CNN) architectures to initialise weights (VGG16, VGG19, ResNet, InceptionNet, etc.).

Multiple researchers have published pre-trained models that aid researcher on avoiding tedious tasks of training data and optimising the cost of running algorithms on hardware. Transfer learning is a new learning framework that allows the use of pre-trained models from other researchers. Dong et al. exploited the transfer learning strategy to train deep CNN due to the lack of age annotated facial images [64]. They state that transfer learning includes pre-train and fine tune where in the former, the randomly initialised networks are first trained with a fair amount of labelled data and in the latter, learned parameters in the mentioned former process are used as an initialisation for a new task.

Well documented pre-trained models for age estimation are communal in the Caffe Model Zoo. In 2015, Levi and Hassner disseminated a deep CNN for age and gender classification [153]. Their model was trained with the Flickr dataset of facial images in the wild [70], to raise performance in learning representations when limited data is available. Training each network required about 4 hours using a robust graphics processing unit (GPU). Similarly, Chen et al. [48] proposed a ranking CNN-based framework for age estimation also trained over the Adience dataset used by Eidinger et al. [70]. Finally, we take into consideration a pre-trained model external to the model zoo but compatible with the Caffe framework. The Deep EXpectation (DEX) model approaches the automated estimation of facial ages with a CNN [218].

Transfer Learning is another optimisation strategy to prevent over-fitting. Knowledge transfer, inductive transfer or transfer learning makes use of existent available data to aid the learning on the new target data, which is composed of training and testing [52]. The use of transfer learning has been increasing throughout the years and has been brought to the attention of researchers where several of them have published pre-trained models to assist other researchers and prevent them from executing the tedious task of training data to solve a specific problem. Inductive transfer can be beneficial when there is lack of labelled data, copyright issues or when data could be easily outdated. Throughout the study, several issues have been encountered. Efforts to attempt to obtain a sufficient quantity of labelled facial age images has been accomplished. However, issues arise due to copyright restrictions, general data protection regulation (GDPR), and ethical concerns. Therefore, a transfer learning solution is required. Transfer learning is usually expressed through the use of pretrained models. Less training data is required when successfully transferring a pretrained model to another task.

### 2.4.6 Regularisation for Deep Learning

#### 2.4.6.1 Dataset Augmentation Techniques

Data augmentation is known as a pre-processing technique in computer vision that attempts to find a data-space solution to the lack of data. It is a convenient method to improve performance and reduce generalisation errors and overfitting. Data augmentation can be achieved through geometric/photometric transformations and synthetic data. The former refers to transformations that alter the initial image, making the CNN invariant to geometric or photometric changes. The latter refers to the use of GANs to create synthetic data that can aid the training process. These networks are further discussed in Section 2.4.6.4.

Image augmentation for facial recognition is not a new topic. It has been studied in the past and has become increasingly popular. Data augmentation can improve the performance of machine learning models and convert bounded datasets into exploitable big data [230]. Lv et al. proposed 5 data augmentation methods: landmark hairstyle, glasses, poses and illumination manipulations. The approach enlarges the training dataset, which aids the impacts of misalignment, pose variance, illumination and occlusion [166]. For facial age estimation, data augmentation was proposed by Liu et al. Their augmentation approach consisted in the application of geometric and photo-metric transformations such as flipping, rotating, scaling, and noise addition. The method aids over-fitting, enhances the robustness of the model and improves the accuracy of age estimation [160].

An ideal underage balanced dataset to solve age estimation as a classification problem would be of size 90k, 5000 images per class as suggested by Goodfellow et al. [90]. Augmented images should be used only on the training dataset. However, test-time data augmentation has been proven to reduce appearance variations and improve face representations [171]. It is convenient to apply data augmentation transformations after the images have been pre-processed.

#### 2.4.6.2 Early Stopping

Early stopping is a DL regularisation technique used to prevent over-fitting and gain the lowest validation error. While training a model, if the error rates do not improve during a certain tolerance (specific number of iterations), the whole training is stopped. Each time the validation error improves, the model parameters are stored. Goodfellow et al. states that early stopping is an efficient hyper-parameter selection algorithm [90]. In several approaches of our age estimation models, early stopping is considered usually with a tolerance proportional to the number of epochs. The ratio is usually a tolerance of 10 to 15 per 100 epochs and 25 to 50 per 1000 epochs. But usually to increase the speed of training, 10 iterations are considered for 100 epochs which equates to a 10% of the epochs.

#### 2.4.6.3 Ensemble Learning

Ensemble learning refers to tackling the ML problem through several models and then averaging them. Usually this regularisation technique produces always better results with higher generalisation performance than the best base model alone [62]. Ensemble methods are usually computationally expensive. Nevertheless, the correct use of the type of ensemble can improve significantly the performance of a single complex model.

Hierarchical ensembles based on the Gabor Fisher classifier [234] and independent component analysis pre-processing techniques [161] are some of the earliest ensembles employed in face recognition. In the CSEM pipeline of designed ensembles discussed in Section 2.2.6.3 an ensemble is used to tackle a CSEM problem. Furthermore, a divide and conquer approach could be introduced to create several ensembles and improve age prediction accuracy. Commonly used ensemble learning algorithms are stacking, bagging and gradient boosting.

#### 2.4.6.4 Generative Adversarial Networks

A GAN consists in two neural networks (Discriminator $D$ and Generator $G$) competing against each other while improving over time, until $G$ achieves such performance where $D$ cannot distinguish that the new generated image is fake anymore (the discriminator model is fooled about half the time), meaning the generator neural network is generating plausible samples. Bowles et al. describe GANs as a means to release information from a specific dataset [31]. Although GANs can be used maliciously to create fake images, fake news, fake videos, etc., the benign applications are significant. GANs are able to estimate images of victims by creating aged versions from an input image [68]. They have also been used for data augmentation by creating artificial instances from a dataset with new produced images that have been introduced randomness but retail similar characteristics of the distribution of the dataset.

### 2.4.7 Evaluation Protocol

The evaluation protocol refers to the steps taken prior to the evaluation of the performance of the model. Several protocols impact in the outcome and hence are documented for revision. Each model that was developed followed the evaluation protocols discussed in this section. The evaluation protocol steps for the dataset are divided into: curation, pre-processing, split and augmentation. **Dataset curation** is analysed in Section 2.3 and **dataset augmentation** is discussed in Section 2.4.6.1.

#### 2.4.7.1 Dataset Pre-Processing

Once the dataset is completed, facial image pre-processing should be applied. Face detection is usually needed for age estimation. While reducing the number of pixels to be evaluated, unwanted background and noise should be also addressed. Many studies convert the images to grey-scale [61, 263], but in our case, skin tone has been proven to mildly influence on the age estimation accuracy [9], and hence the colours should be preserved.

#### 2.4.7.2 Dataset Split

As an evaluation protocol, each proposed model has to be trained, validated and tested on a dataset that has been split accordingly. The ideal split would be a 80/20 split for training and validation, and a 80/20 for validation and testing respectively. But this might be an issue with small sized datasets. With less training data, the parameter estimates have greater variance. A test dataset is considered optional but paramount to evaluate the final performance of the model fit on the training dataset. A validation dataset is used to automatically select the best classifier during the training.

### 2.4.8 Network Architecture

The relevant network architecture should be employed depending on the type of problem to be solved (regression or classification). Simple neural networks can be used when the inputs are specific and limited vectors represent the image; i.e., facial embeddings can produce 128 or 512 facial vector representations via FaceNet (As explained in Section 3.2.5). Much more complex networks such as VGG16, VGG19, ResNet50 and Inception are CNNs that have been trained on the ImageNet Dataset and can be used for age estimation classification of raw facial images as inputs. Figure 2.7 depicts a VGG16 architecture with facial images as inputs, the large squares represent the convolutional + rectified linear unit (ReLU) layer and the adjacent smaller red squares represent max pooling. The last 3 layers are fully-connected (FC) layers and the last layer is also a softmax of *K* outputs. For underage age estimation, ideally *K=19* neurons should be considered. From newborn, 1 year old to 18 years old. Nevertheless, due to the average MAE in the literature being 4.15 (Average calculated from Table 3.2), the upper bound age limit should be set to the age of an adult (18 years as discussed in Section 2.2.1) plus the 4 years of error equating to 22 year-old subjects.

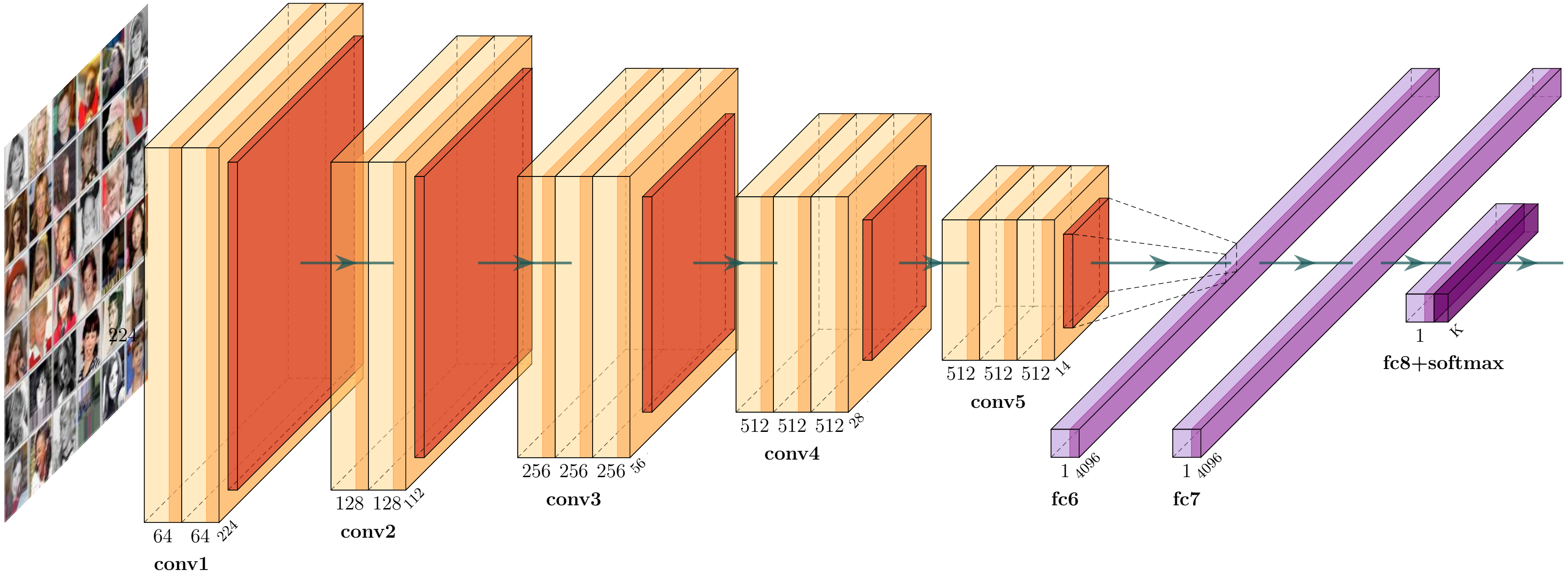


Figure 2.7: VGG16 Macro-Architecture with Facial Images as Inputs and *K* Outputs in the Softmax Layer.

### 2.4.9 Hyper-Parameters

Selecting the adequate hyper-parameters is a complex task. Each ML problem has its different complexities to solve and settings that are used to control the behaviour of the learning algorithm [90]. The hyper-parameters are usually updated accordingly in the validation set. The validation set represents typically 20% of the dataset. Whereas the training set equates to the rest (80%). These values are arbitrary but have been influenced by the Pareto principle.

Hyper-parameter tuning is referred to searching the hyperparameter space for the optimum values. It may be accomplished either manually or with an optimisation framework such as SMAC, Spearmint, Hyperopt, GPyOpt,Vizier, Katib, Tune, Autotune, Optuna, etc. [3]. Optuna is framework agnostic and can be used with any ML or DL framework such as Theano, Pytorch, Tensorflow, Keras, Scikit-learn, etc. The advantage of using Optuna over a manual approach is that the experiments can be integrated in a single ML logging framework. Neptune is a light-weight management tool that keeps track of ML experiments [186]. The tool enables live monitoring, has integration with Jupyter Notebooks, allows the search and comparison of experiments and entitles collaboration with other users.

The 3 steps to optimise a problem are the following:

1. Define an objective function to be optimised.
2. Suggest hyper-parameter values using trial objects.
3. Create a study object and invoke the optimisation method over $n$ trials.

#### 2.4.9.1 Learning Rate

The learning rate (LR) is one of the most important hyper-parameters for deep neural networks (DNNs). It refers to the size of the step for each iteration, while approaching the minimum of a loss function [182]. The value is usually a positive scalar between 0.0 and 1.0. The LR value is inversely proportional to the training time. The lower the value, the more resources it would require to train.

#### 2.4.9.2 Optimisation

Optimisation is one of the main components of machine learning. Gradient descent is an optimisation technique used to find the minimum of a function. It is regularly used in deep learning models to update the weights of a neural network. The following gradient-descent-based optimisation algorithms have been used in our studies to lessen the error rates:

- adaptive moment estimation (ADAM) [138]
- adaptive gradient (ADAGRAD) [69]
- stochastic gradient descent (SGD) [30]
- stochastic weight averaging (SWA) [118]

ADAM and SGD are commonly used to optimise deep neural networks and are widely used in age estimation. For our research, we explore the use of several gradient-based optimisers and focus specifically on the novel SWA. SWA is a procedure that enhances generalisation in DL models over SGD at no additional cost. Izmailov et al. proved that the SWA procedure is able to find much flatter solutions than SGD and the solutions are wider than the optima found by SGD [118]. The authors also notice an improvement in the test accuracy versus SGD training on several state-of-the-art residual networks. It also has slightly worse train loss, but better test error.

### 2.4.10 Artificial Intelligence for Digital Forensics

Lack of time and experts influence negatively on a crime investigation. To process huge amounts of data in a timely manner and enhance digital forensic investigations, multiple AI approaches have been proposed. The use of automation in DF has been highly criticised due to the constant need of a human-in-the-loop approach. A human check is still required to not miss any critical information with probative value. Furthermore, push-button forensics besides having technological challenges, have also political and social implications [121].

Automated prioritisation and evidence discovery practices in DF could help sift through irrelevant data/devices and control the workload to a more manageable state. In 2013, Marturana and Tacconi used triage to categorise digital media [170]. Their study evaluates the use of popular mining algorithms for tackling two cases studies: copyright infringement and exchange of child abuse material.

A CNN is a deep artificial neural network (ANN)[18] that is capable of classifying and clustering visually similar images, and recognising objects in a scene. CNNs have been

[18] Computer systems that learns to accomplish tasks by observing examples rather than executing a specific algorithm [255].

brought to attention in various fields for their promising results. DF has also been influenced by the attractive qualities of CNNs.

In image forensics two problems are addressed: image tampering & image sourcing. For the former, multiple studies have been conducted with deep learning based techniques [21, 46, 208, 269]. Three studies employ CNNs and one of them implements a Stacked Auto-encoder(SAE) approach. For the latter problem, identification of the acquisition device that produced the data could link the electronic device to the perpetrator. Tsai et al. were able to obtain highly accurate predictions with SVM on similar photographed scenes generated both by traditional and mobile-phone cameras [240].

Not until recently, studies involving multimedia forensics (branch of DF that studies collected multimedia signals such as audio, video and images from mass storage devices) and DL have flourished. CNNs have shown promising results on image recognition, video analysis and natural language processing(NLP). Freire-Obregon et al. [81] have implemented a source camera identification (SCI) method that is able to infer the noise pattern of mobile camera sensors/fingerprints.

### 2.4.11 Object and Face Recognition

Object detection is one of the core problems in computer vision that deals with identifying and locating instances of objects of certain classes present in digital visual media. For a human, it would take a glance to detect an object. Conversely, for a machine it would require extra effort to replicate the intelligence of individuals. This is the main goal of object detection.

Face and landmark detection are well-researched domains of object detection. These techniques typically leverage ML or DL (depending on the data and processing capacity), to produce relevant results. Facial recognition is a data-driven technology that has become popular across the globe and significant advancements have been made in recent years. An exponential increase on users has been observed in common everyday applications. Moreover, biometric systems are expanding their robustness with the addition of facial-based authentication factors that prevent impersonation attacks, e.g., Apple's Face ID and Android's face recognition technologies. Facial recognition as a biometric system, is a widely-used technology that maps facial features from images to detect faces and recognise the associated identity. For further information relevant to facial recognition, refer to Section 3.1.

CHAPTER

THREE

# RELATED WORK

## 3.1 Face Recognition

### 3.1.1 Introduction

Facial Recognition applications have been commonly found in airports, mobile devices and certain web pages [82]. Facial recognition is shaping the future of several security innovations: facial security checks could be used to prevent credit card cloning, smartphone unauthorised access, fraudulent exam takers, fake social media accounts, etc.

Images of faces are easy to obtain regardless if they are produced from a modern camera or are collected from the internet. Challenges arise when the expected images are not in controlled environments such as the ones produced in visa applications or mugshots [123]. Other external factors such as hair, makeup, jewellery, glasses, head-wear, etc., will all influence the appearance of the face and will have an impact consequently on facial age estimation as explored in Section 3.2.11.

Dimensionality reduction of redundant sampling is one of the main problems with face recognition technologies [123]. Principal component analysis (PCA) is a statistical technique used often for face recognition and that can extract the so-called "eigenfaces" which are significant statistical information related to the variance of faces in a set of face images.

### 3.1.2 Facial Features

Facial features are composed by multiple landmarks. Each landmark represent an identifiable point present in a face. Facial landmarks serve as anchor points on a face graph [42]. In Figure 3.1, hundreds of facial landmark points can be seen depicted.

These points have been detected by the Face++ application programming interface (API) which is discussed in Section 3.1.3.2. Commonly used points are the eye corners, the nose tip, the nostril corners, the mouth corners, the end points of the eyebrow arcs, ear lobes (weather attached or detached), chin (weather squared, round or pointy), etc. Attributes such as eyes (round, almond, round-almond), nose shape (straight, round, wide), eye-browns (thin, medium, thick), mouth, jawline, etc., are shapes that are conformed by several joint landmark points. The main issue of automatic feature-finding algorithms is that in a low-resolution scenario, identification and positioning accuracy is compromised [123].

For missing children investigations or in the identification of suspects, a feature that is vastly important is the smile which is a facial expression that witnesses often see and recognise.

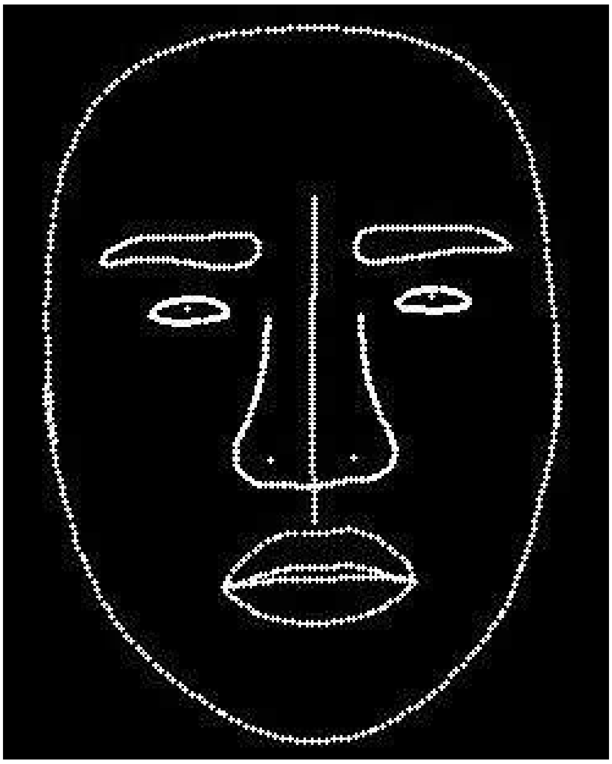


Figure 3.1: 1000 Facial Landmarks Captured using Face++ Application Programming Interface

A decade ago, automatic and accurate facial feature detection seemed a difficult problem due to the uniqueness and variety of human faces, expressions, facial hair, poses, glasses, lightning conditions, etc. [176].

The approach presented by Milborrow and Nicolls in 2008 was based on active shape models (ASMs) which were first introduced by Cootes et al. [50]. The ASMs were used to locate features in frontal views of faces that were in an upright position. In the same year, Horng et al. studied age group classification based on facial features. Their approach was limited to gray-scale facial images with 4 age groups. A Sobel edge operator and region labelling was used to obtain the positions of the eyes, nose and mouth [110]. It is stated by Kwon and da Vitoria Lobo that for an age classification task facial features suffice [146].

### 3.1.3 Face Detection

Face detection and alignment are essential to many face applications such as facial recognition, expression analysis and age estimation. However, factors such as image resolution, occlusion, brightness, contrast, roll/yaw/pitch variations, etc., present challenges in real world applications.

Edge detection plays a key role in face recognition. Sobel edge operators were proposed in the past to identify and extract successfully edges in object recognition. Edge maps aid the representation of faces as a single unit [246]. Martínez et al. proposed Sobel edge detection for face recognition, used in the pre-processing phase on the training data to produce better performance in the modular neural network [169].

Viola and Jones proposed a cascade face detector that uses digital image features (Haar-like features) and AdaBoost[1] to train cascaded classifiers, while achieving good performance and real-time efficiency [268].

Age estimation as well as face recognition, requires in some cases facial detection for the pre-processing phase. Elimination of background noise and unrelated features to the face can aid the performance of the age estimation model. In 2016, Antipov et al. used a rigid-template-based facial detector with a multi-view facial landmark detection tool for alignment. The error rate achieved for their age estimation model was a validation of 0.2609, a metric $e$ defined as the size of the tail of the normal distribution with the mean $m$ and the standard deviation $s$ with respect to the predicted value $x$.

#### 3.1.3.1 Offline

Several offline face detection methods have been built. The most famous open source libraries are `dlib` and multitask cascaded convolutional network (MTCNN). The former is a C++ toolkit containing machine learning algorithms. The `dlib` toolkit contains face detection based on the classic histogram oriented gradient (HOG) feature descriptor combined with a linear classifier, an image pyramid, and a sliding window

[1] AdaBoost is an ensemble learning method which was created initially to improve the efficiency of binary classifiers.

detection scheme [137], and is the most used tool for effectively and swiftly detecting fiducial points with traditional computers in a few milliseconds [133]. Leong was also able to accomplish a high speed detection of 30 millisecond on a mobile application that implemented DL of facial embeddings and facial landmark points for the detection of academic emotions [151]. Furthermore, when no face is detected (despite being visibly present), the CNN version of `dlib` may be executed. This technique would lessen the processing time and increases the face recognition hits. The latter face detection method works on three stages and uses a single neural network for each process. Moreover, it employs a deep cascaded multitask framework that exploits intrinsic association between alignment and detection to increase performance [268].

The main advantage of using offline face detectors are that they are free and open source. It is possible to fine-tune the model and also run unlimited facial detection's without paying additional costs to cloud-based servers. Additionally, storing personal identifiable information on an unknown server is a great risk.

#### 3.1.3.2 Online

Online cloud-based face detection services have been recently developed. They usually are part of an AI face functionally and have an established cost. If the cloud service provider (CSP) delivers facial recognition or facial age estimation services, a face detection component may also be offered.

**Face++**[2], a Chinese cognitive service is capable of several facial recognition tasks such as face detection, face comparison and face searching. It has been bench-marked by Jaeger et al. against other cloud services with promising accuracy [119]. Also used in many variety of research from face detection, gender/race to facial age estimation [10, 113, 211].

**Kairos** is a Greek word meaning the right, critical, or opportune moment. In cyberspace, it is a face recognition company that provides several facial feature applications such as face identification, face verification, face/age/gender/multi-face/landmark detection, diversity recognition, etc. Many of the products have been mentioned or used in research [7, 58, 206].

Amazon, Google, IBM, Microsoft Azure, Huawei Cloud as CSPs, all have face detection capabilities and provide an API for integration. The main advantage of using online cloud-based face detectors is the speed obtained by each request. Top-edge machines that are offered by cloud service providers are powerful but require a cost per

[2] `https://www.faceplusplus.com/`

transaction. However, in dealing with CSEM investigations, LEAs cannot transmit this sensitive information to a third party service.

## 3.2 Soft Biometrics

### 3.2.1 Introduction

Facial recognition is a well-known topic studied across several fields. The race to obtain a highly accurate tool to recognise, verify and cluster similar faces is a common task within the topic of computer vision. The deluge of facial photographs on the cloud has allowed the creation of robust facial recognition systems. Nevertheless, labels for soft biometric traits such as age, gender, ethnicity, height, weight, eye colour, marks, etc. are scarce. While the consideration of these traits are able to improve the accuracy of a biometric system [120], there are rarely datasets that contain such information accurately annotated. Age is a cue for face verification and facial recognition widely used in forensics. Automated age prediction could be valuable as an aid to live and post-mortem triage, both collected and stored evidence while assisting digital forensics backlogs. Age prediction can also assist in the identification of victims or suspects in CCTV footage, photographs, or CSEM. Moreover, GANs are able to estimate images of victims by creating aged versions from an input image [68].

Age estimation models rely on good quality images with the relevant age labels. Nonetheless, accurate age annotations in facial datasets are also inadequate. Certain age groups have less amount of data, particularly the underage age range. Datasets for this age range are difficult to find due to certain restrictions and ethical implications.

As mentioned in Section 2.1.2.7, an exponential accumulation of devices in digital forensic laboratories is an ongoing issue that has contributed to backlogs in LEAs over the past years throughout the globe [225]. Intelligent automation is needed to expedite digital investigations that are hampered by lack of resources, such as time and skilled expertise. Moreover, Sanchez et al. [223] verified that digital forensic practitioners demand automated tools to detect CSEM, age estimation and skin tone detection and intelligent artefact prioritisation can expedite digital investigation [66]. The automation of some of those tools are presented in this section.

### 3.2.2 Age Estimation

Determining an individual's age can bring a forensic investigation one step closer towards the identification [2]. Different people age differently. Even a single year can make a significant difference in the appearance of a child. There are several factors to identify a missing person: gender, ethnicity, parents and how parents/grandparents looked like at the age of the missing subject, siblings, what do they look like, their lifestyle, weather the person practised a sport, exercise, diet, health state when the person was last seen, vices such as smoking, drug abuse, etc.

The human face can reveal important information, such as gender, approximate age, skin tone, race/ethnicity, eye colour, hair colour, presence/absence of makeup, presence/absence of beard, presence/absence of moustache, etc. All these elements are know as soft biometric traits. Dantcheva et al. [55] defines soft biometric traits as "physical, behavioural or adhered human characteristics, classifiable in predefined human compliant categories".

Accurately determining the age of a victim can prove crucial in a CSEM possession and/or distribution case, especially for borderline age ranges between underage teenagers and young adults. The prediction of age as a soft biometric trait has been proven to be difficult due to the absence of strong cues that determine the oldness of a subject. Kloess et al. suggest that discrepancies between the face and body, natural variation between different ethnicities and the environment that the person is exposed to are factors that affect the age prediction process. The aforementioned research takes into account multiple factors that can lead to the classification of an image either if it is an indecent image of children and the respective age group.

### 3.2.3 Human Facial Age Perception

Humans are quite accurate at estimating the age of other humans. The error rate has been measured to vary from as low as 2.07 years and as high as 8.62 years depending on a variety of factors including the age of the assessor, the age of the subject, and the difference between both [179]. The age of young people tends to be consistently overestimated [105, 202, 260] and a tendency to assimilate the estimated age with one's own age is suggested [245, 248]. Moyse and Brédart [179] presented a study on own-age bias in the accurate estimation of faces. The authors found that their 114 participants were more accurate at estimating the ages of those within their own age-group (10-14, 20-30, and 65-75 years old). The accuracy of human age estimation of others can also be negatively impacted by a range of other factors including gender [248] and emotion/facial expressions [85, 248]. Neutral expressions results in the highest accuracy, whereas any other expression results in less accurate estimations [248].

Perhaps the most relevant studies to the borderline adulthood focus of the work presented as part of this research are those focused on age estimation for the sale of age-restricted products, such as alcohol and tobacco. In 2001, Willner and Rowe [260] measured the accuracy of servers of alcohol in age estimation for males and females aged 13, 16, 20 and 22 years old. Their results showed that the estimated ages of teenagers was often overestimated. A 3% of 13 year-old males and 18% of 13 year-old females were judged to be of legal drinking age in the UK (18 years old), rising to 38% for 16 year-old males and 56% for 16 year-old females. This is consistent with the results presented by Voelkle et al. [248]. In their work it was found that those working in the sales of age restricted products had a mean accuracy of 3.26 years for the 15-19 years age range, versus a mean accuracy of 4.01 years for a control group, i.e., those working in roles not relying on age estimation as part of their daily duties.

### 3.2.4 Facial Age Estimation

Automated facial age estimation requires three main properties: validated facial age labelled datasets, robust face detection, and ML. Many existing approaches reuse the same dataset(s) either for bench-marking, validation, training or testing purposes.

#### 3.2.4.1 Offline Models

In the early stages, facial anthropometric[3] models were suggested for age prediction. In 2006, Ramanathan and Chellappa [207] proposed a cranio-facial growth model that classifies growth related shape variations observed in underage faces. Their model was capable of face recognition across age progression. The dataset for the aforementioned model was the FG-Net ageing dataset (82 subjects with ages ranging from 0 to 69 years old and over 50% juvenile subjects[4]) and a separate dataset containing 233 images [207].

In 2009, Guo et al. [98] found that age estimation performance was able to improve when manifold learning uses biologically inspired features (BIF). The accuracy achieved is positively influenced by a known gender. Therefore, their approach consisted in two different MAEs for each gender. Furthermore, the model was a combination of "BIF locality sensitive discriminant analysis" and "BIF marginal fisher ana-

[3] The science of measuring sizes and proportions on human faces [207]

[4] https://yanweifu.github.io/FG_NET_data/index.html

lysis", reaching MAE rates of 2.58 years for males and 2.61 years for females. The data used was the large Yamaha gender and age (YGA) database that contains 8,000 outdoor facial images of Asian subjects. The dataset was distributed equally by gender and divided in age ranges from 0 to 93 with intervals of 9 classes per age group until the age of 70 and a group containing the rest due to the lack of images available for subjects over 70.

Eidinger et al. presented an approach on age and gender estimation using standard linear SVM with their own dropout-SVM scheme. The dataset employed was Adience, which is a collection of CC images sourced from Flickr. Furthermore, the ages were labelled in heterogeneous categories in an unknown manner that contained in average 2,205 images per age group, but were unbalanced.

More recently in 2018, Rothe et al. [218] proposed a DL solution based on a VGG16 CNN architecture pre-trained on ImageNet[5]. This achieved a MAE of 3.252 years. The data was trained on a dataset named IMDB-WIKI, consisting in over half a million facial images of celebrities (currently the largest public face image dataset annotated with age and gender labels) that had been crawled from IMDB and cross referenced with the age denoted in Wikipedia.

Geng et al. reiterate the lack of sufficient and complete training data [87]. However, the authors exploit the fact that close ages look quite similar. Instead of labelling with a single age, a label distribution is considered. Small datasets were used: FG-NET which is an ageing dataset of 1,002 subjects [147], and MORPH which is a larger dataset of over 55k images [212]. The best performing results in terms of MAE oscillate between 4.76 and 8.06 years, The MAE in different age ranges was also evaluated in the FG-NET dataset. The best performance lies on the age range 0 to 9 (2.30 years) followed by the age range 10 to 19 (3.83 years).

Chao et al. aimed to overcome the data imbalance problem related to the number of images per age. An imbalance treatment is introduced to the training phase and the connections between facial features and age labels by combining distance metric adjustment and dimensionality reduction, are explored. Performance evaluated on the most widely-used FG-NET ageing database produced MAEs ranging from 3.06 and 3.10 years for ages smaller than 30. The MAE in different age ranges were also evaluated with the aforementioned database. The best performance lies within the age range 0 to 9 (1.911 years) followed by the age range 10 to 19 (3.52 years) with the C-lsLPP algorithm approach.

In 2017, Liu et al. presented a Group-aware deep feature learning approach that con-

[5] `http://www.image-net.org/`

sisted in learning a discriminative feature descriptor per image of the raw pixels for face representation [158]. The main motivation is that age labels are chronologically correlated and face ageing datasets lack labelled data in certain groups. The datasets used were FG-NET, MORPH and the Chalearn Challenge Dataset [71]. The corresponding MAEs are 3.93, 3.25 and 4.21 years respectively.

#### 3.2.4.2 Online Models

The main advantage of using cloud-based biometric services is that the results obtained are processed by state-of-the-art classifiers developed by experienced companies in the space, such as Amazon, Microsoft, and IBM. The main disadvantage of online tools is the ongoing costs associated with their usage. Most of the service responses are configured in JavaScript Object Notation (JSON), which allows an easy integration with the performance evaluation.

In 2010, Amazon acquired “Rekognition” from an AI start-up company, Orbeus [205]. The company had developed a facial recognition software that detected traits on images with ANNs. ANNs are systems that learn to accomplish tasks by observing examples rather than executing a specific algorithm. They are structured by an initial input layer of neurons, one or more hidden layers, and a final layer of output neurons. Machine learning as a service (MLaaS) was introduced to facilitate non-experts in the training of models without expertise in the topic. The Rekognition service is a DL-based image analyser that is able to detect age with a minimum and maximum value as a dual class output. The most suitable results between the dual class outputs and the mean value were assessed. The investigation led to the use of the minimum value, which is also a good practice for the procedures in a digital forensic case where the cost of inaccuracy is potentially high. Kairos (a previously known free online service) uses a SVM algorithm for the model to help isolate different types of faces into the corresponding age class. Nevertheless, the performance is low compared to the rest of the age estimation services [7]. Finally, Microsoft Azure Cognitive Service uses a Multi-layered deep learning methodology [258].

### 3.2.5 Facial Vector Embeddings and Soft Biometric Traits

Face embeddings are high-grade features extracted usually from detected faces. They use deep convolutional neural networks (DCNNs) to map a facial image to a vector. The most used model is FaceNet which predicts features that are an array of 512 vector representations. The model is trained through a triplet loss function that influences facial embeddings for the same subject to have smaller distances and different subjects to have larger distances [228]. In this research, the cosine similarity is used to calculate with a given threshold, if the facial embedding arrays belong to the same identity in a given multidimensional space. Equation 3.1 is the cosine similarity between face *a* and face *b*, where the value of *n* is 512.

$$\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} = \frac{\sum_{i=1}^{n} \mathbf{a}_i \mathbf{b}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{a}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{b}_i)^2}} \tag{3.1}$$

Schroff et al. [228] proposed a system that learns mapping from facial images where the distance between the vectors produced are able to determine facial recognition, verification and clustering of similar images. The output creates embeddings of 128 dimensions per face but currently 512 dimensions are supported. Face embeddings refers to the facial features that can be extracted from a facial image. Once processed, the problem becomes a k-nearest neighbours (KNN) classification problem.

Work accomplished with facial vector embeddings for trait related research such as age, gender, emotions and attractiveness has only been recently exploited in the past two years. In 2018, Jekel and Haftka used a Logistic Regression and a SVM approach to automatically review online dating profiles based on the user's historical preferences [125]. The authors discussed a possibility of the Facenet vectors being related to attractiveness. This research was one of the first using Facenet Face Embeddings for tasks other than facial recognition.

Later in 2019, Terhörst et al. proposed a multi-algorithmic fusion for age and gender estimation based on stochastic forward passes through a dropout-reduced neural network ensemble [237]. Their approach was benchmarked on the Adience dataset [70], and achieved an age estimation accuracy of (64.6 ± 2.8).

Recently in 2020, Swaminathan et al. developed a method to predict gender based on several machine learning classification techniques on Facial Embeddings. Logistic Regression, SVM, KNN, Naive-Bayes and Decision Trees where evaluated on the UTK Face Dataset [270] and the best performer KNN achieved an accuracy of 97%. In the same year, facial embeddings and facial landmark points for the detection of academic emotions such as engagement, frustration, confusion and boredom, were studied by Leong [151]. The author evaluated the use of deep learning on FaceNet embeddings and facial landmark points and hypothesised that the facial embeddings may similarly offer valuable information for the detection of emotions. A Long Short Term Memory (LSTM) network architecture was used and the accuracy to detect both boredom and frustration was 52.15 and 70.67 % respectively.

### 3.2.6 Regression vs Classification

In our daily life, age is treated as a discrete variable, except among the very young - a five and a half year old will not be denied their half year. Ages are binned more coarsely as we age - young, middle-aged, old - and much previous work has treated the age estimation as a classification problem, with samples assigned to broad age buckets. In the extreme, it is binarised: minor or adult? When a more accurate age estimate is wanted - and accurate ground-truth is available - we can instead treat age as a continuous variable, and make age estimation into a regression problem. Regression structures are used to estimate a value (continuous inputs) instead of a fixed class, leading to an infinite set of possible outcomes [28].

Whether modelled as regression or classification, data is, as with all machine learning problems, the limiting factor. Age estimation has been addressed in the past with several machine learning regression techniques; predominantly Support Vector Regression (SVR), Random Forests (RF) and Canonical Correlation Analysis (CCA) [78]. Conversely, commonly used classification algorithms such as KNN, multilayer perceptron (MLP), AdaBoost and SVM have been studied to perform accurate age prediction and grouping [155].

The difficulty with using fine bin sizes (i.e., 1 year wide) while also taking a classification approach is that the model will score the same loss for being wrong by 1 year as it would for being wrong by 20 years. With regression, a large error (e.g., 20 years) can induce a larger weight update than a small one (e.g., 1 year). The difference would matter less if a binary classifier was trained, but for samples close to the decision boundary it would still matter.

### 3.2.7 Underage Facial Age Estimation

Work specialised in underage facial age estimation has been limited due to the challenges of collecting data which is understandably subject to ethical implications, lack of underage datasets, and scarcity of reliably annotated images. Nevertheless, apparent age estimation on children was studied by Antipov et al. in 2016. Antipov et al. used a fine-tuned VGG16 (a very deep CNN of 16 weight layers used for large-scale image classification) to train a model of minors; they documented their winning approach for the ChaLearn LAP competition on apparent age estimation; since the major challenge was the age estimation of children, the authors created a separate VGG16 model for minors from 0 to 12 years old and integrated the model to the final solution. HeadHunter, a detector based on rigid templates, was their choice for face detection and the alignment technique was based on a multi-view facial landmark detection tool. The error rate achieved for validation was 0.2609, a metric $\epsilon$ defined as the size of the tail of the normal distribution with the mean $\mu$ and the standard deviation $\sigma$ with respect to the predicted value $\hat{x}$. In the same year, Ferguson and Wilkinson [77] determined that manual human visual age estimation of children's faces reveals poor accuracy, confirming the difficultly to precisely predict age. They also suggested that black and white images were classified with less accuracy.

Age estimation classification models tend to perform better when the age bins have been grouped; hence the number of classes decreased. The penalisation for wrong classifications are less severe; similar to the effect of using a regression-based model vs a multi-class classification model. Furthermore, limiting the size of the evaluated age range to underage subjects can potentially create an easier problem for age estimation. If the complexity of the problem is gradually increased, this is known as CL where the speed of convergence of the training process is increased [23]. Refer to Section 2.4.5.1 for more information about CL.

### 3.2.8 Gender Estimation

Gender has been predicted as a binary classification problem in the past and has received considerable attention over the past years. Humans are able to distinguish between a male and female, achieving accuracy above 95% based on looking only at the face [34]. Gender classification is paramount as it can enhance the performance of face recognition and human-computer interfaces [254]. Early approaches to automatically classify gender with neural networks were proposed in 1990 [89]. A novel approach to detect gender based on the colour of human faces was proposed by Nestor and Tarr [187].

### 3.2.9 Skin Detection and Skin Tone Classifiers

To lessen the exposure to CAM, multiple approaches have been considered. Skin detection algorithms could potentially sift unnecessary images and flag inappropriate content. In 2005, Ap-Apid[15] developed a skin color distribution model based on RGB. His nudity detection algorithm had a 95% recall with a 5% false positive rate. Later in 2016, Deep CNNs were used by Nian et al. [188]. The latest demonstrates the advantage of using AI over hand-engineered visual features that are hard to analyse and select. The notable trend of CNNs has been flooding research topics in the past years.

Automated detection of skin tone has received considerable attention from researchers – specifically for biometrics and computer vision applications [136, 168]. The impact of two approaches has been evaluated: simple skin detection (SSKD) and face colour extraction (FCE). Both approaches are based on k-means clustering[6] in order to determine and classify a subject's skin tone.

SSKD refers to unsupervised skin tone estimation/segmentation. The approach predicts skin tone from an image of a subject, while doing a rough segmentation of the skin based on a pixel-wise classifier [266]. The algorithm consists of two main components: foreground/background separation using Otsu's Binarisation and pixel-wise skin classifier based on hue, saturation, value (HSV) and YCbCr[7] colour spaces [37]. The FCE approach initially detects the facial landmarks using the `dlib` library [137]. Subsequently, noise is removed by applying the convex hull algorithm[8] on the facial land-marked point. Finally, the RGB values of the skin are computed using a histogram-based clustering algorithm.

### 3.2.10 Race/Ethnicity Prediction

Lu et al. [164] addressed the ethnicity identification from facial images in a machine learning framework. They employed a Linear Discriminant Analysis (LDA) based scheme for the input face images at multiple scales, to classify Asians from non-Asians. The dataset used was balanced between Asians and non-Asians of 2,630 faces. Their LDA Ensemble approach produced an accuracy of 96.0% for Asian subjects and 96.6% for non-Asians. The authors argued that ethnicity classifiers do not have to be perfect to be useful in practice. In contrast, Hosoi et al. [111] addressed a novel approach for ethnicity classification with two technologies: Gabor Wavelets Transformation and retina sampling for facial feature extraction, and SVM for ethnicity classification. The authors classified individuals into three major groups: Asian (Mongoloid), European (Caucasian) and African (Negroid). The dataset used consisted in 1,991 images. The dataset was not equally distributed, as there were more images for evaluation for the first two groups. Nevertheless, the ethnicity estimation results demonstrated that despite the lack of images for African subjects, the model was able to classify the images with high accuracy: European: 93.1%, African: 94.3% and Asian: 96.3%.

[6] k-means clustering is a method for vector quantization – mainly used for cluster analysis in the data mining field.

[7] YCbCr is a family of colour spaces used as a part of the colour image pipeline in visual media systems.

[8] Convex hull is a fundamental structure for both mathematics and computational geometry [20]

### 3.2.11 Influencing Factors on Age Estimation

The factors affecting facial ageing have been categorised into intrinsic and extrinsic components [13]. For the former, there are internal factors such as size of the bone, genetics or facial changes due to the development of a child. For the latter, any presence of external factors including the environment, habits, diet, makeup and cosmetics, race, ethnicity, etc.

#### 3.2.11.1 Facial Expressions

One example of influencing factors in age estimation is facial expressions. Voelkle et al. found that happy facial expressions are mostly underestimated [248]; whereas, smiling, frowning, surprise and laughing may introduce facial lines that are confused for wrinkles and thus impact on the age estimation performance.

#### 3.2.11.2 Noise

Noise introduces more error onto the estimation depending on its magnitude. It is a randomness that affects an image due to either brightness, colour or digital encoding, and often occurs during image capture, digital sharing, etc. [76]. The presence of noise in an image is expected to be linearly correlated with performance.

#### 3.2.11.3 Makeup

Facial cosmetics have been found to influence perceived facial age estimation; a simple cosmetic alteration is capable of compromising the outcome of a biometric system [54]. Lip makeup was found to be the most prominent of the cosmetic range with a mild correlation to the decay in age estimation accuracy for specific ages. Moreover, Chen et al. [45] found that the presence of cosmetics can hide facial imperfections caused by age, e.g., wrinkles and dark spots, resulting in underestimation.

#### 3.2.11.4 Gender

The influence of gender on estimation was studied previously where it was shown that the errors exhibit a higher rate of error for female subjects than for males [7]. From this, we can deduce that gender is a soft biometric trait that significantly impacts the overall accuracy of the age prediction model.

#### 3.2.11.5 Ethnicity

Per Moyse [180], little interest has been shown on the influence of race/ethnicity on age estimation. Instead, the occurrence of own-race bias (Similar to own-age bias discussed in Section 3.2.3) has been encountered. This bias was found within Caucasian participants in contrast to African participants that performed equally for both Caucasian and African faces in a study driven by Dehon and Brédart [59]. The authors further suggested that the no own-race bias occurred due to those participants being more in contact with Caucasian faces and hence increasing their expertise for such faces.

Certain studies have determined that there is an influence of ethnicity on age estimation and have shown that the ageing process varies significantly among different ethnicities [97, 99, 227, 256]. Guo and Mu [99] proposed an age estimation framework to reduce the influence of ethnicity and gender. The framework consisted in segregating models per ethnicity and per gender. The performance of the models were reported only for Black and White races due to the samples of other races being too small. According to Schmeling et al. [227], in traditional age estimation forensics, ethnicity has no remarkable influence on skeletal maturation for the relevant age group, and ossification rates depend mainly on the population's socio-economic position. Hence, they concluded that forensic age estimates should consider ethnicity and socio-economic status appropriately.

The influence of ethnicity in age estimation has presented a challenge for both supervised and unsupervised facial age estimation. To "alleviate" this influence, some studies have used a limited version of certain datasets. For example: in [43, 47, 257], the authors have selected a subset of the MORPH dataset (discussed in Section 3.3.4.2) restraining the data to 5,475 Caucasian people. Although this practice of removing certain ethnicities due to the aforementioned influence has apparently become commonplace and accepted in the community, there are major ethical concerns that affect minorities or disadvantaged groups due to the stimulation of algorithmic injustice. These kind of practices not only are inadequate but can also be harmful to vulnerable groups as discussed by Birhane [25]. The author claims that in the process of data cleanse, rich information is stripped away. Furthermore, whilst it is important to have an implementable tangible, scientists around the world should be rethinking and changing their habits on how the data is collected in order to include a more diverse dataset.

## 3.3 Datasets for Facial Age Estimation

### 3.3.1 Introduction

Databases for known sex offenders are quite useful. Combined DNA Index System (CODIS) has been used in the past and many homicides and missing persons cases have been solved with these type of databases [123]; but what happens when there is not a DNA component and only an image is available?.

The importance of large sample size, good quality annotated datasets, and the sharing thereof, with the research community is fundamental. Facial image datasets annotated with both age and gender are needed to assist in solving cases where a photograph or CCTV video has become a crime scene and to train machine learning models to predict further information from incoming data. Grajeda et al. stresses the benefit of sharing datasets within the research community in favour of replicating results [95].

High-quality large-sample-sized facial image datasets annotated with both age and gender are needed to train models that are capable of predicting accurate age. Several age annotated datasets have been released but with certain limitations, such as lack of images in certain age groups, presence of noise in photos that reduce the quality of the dataset, inaccurate age labelling, etc.

Private datasets not only hinder benchmarking but can present wasteful resources that are a result of overlapping research. However, the reasons they may remain private could be due to privacy, ethical and moral concerns. Following in this section, we discuss about the public datasets that have been classified by size: small correspond to the range 10 to 15K, medium correspond from 10k to 20k and finally, the large classification corresponds to over 20k images.

### 3.3.2 Small Datasets

#### 3.3.2.1 FG-NET

FG-NET is a public domain dataset of 1002 images. It is classified as a small dataset and has been used in hundreds of research. In 2013, Han et al. evaluated several of their algorithms with the FG-NET dataset. They found out that most photos of such dataset are dominated by children. This is a motivating fact for our study. However, the numbers shared by FG-NET for underage images were less than our expectations. The FG-NET [253] dataset contains 82 subjects with photographs of each at varying ages ranging from newborn to 69 years old. Although over 50% of images in the FG-NET dataset are child images, the demand for underage training and test data has led to the creation of alternative databases. The distribution of male and female subjects per groups of 10 years can be seen depicted in Figure 3.2. There are predominately more male images than females per age but the distribution is proportionally quite similar besides the first bin (0 to 9 years), where there is a peak of male images surpassing 200 counts.

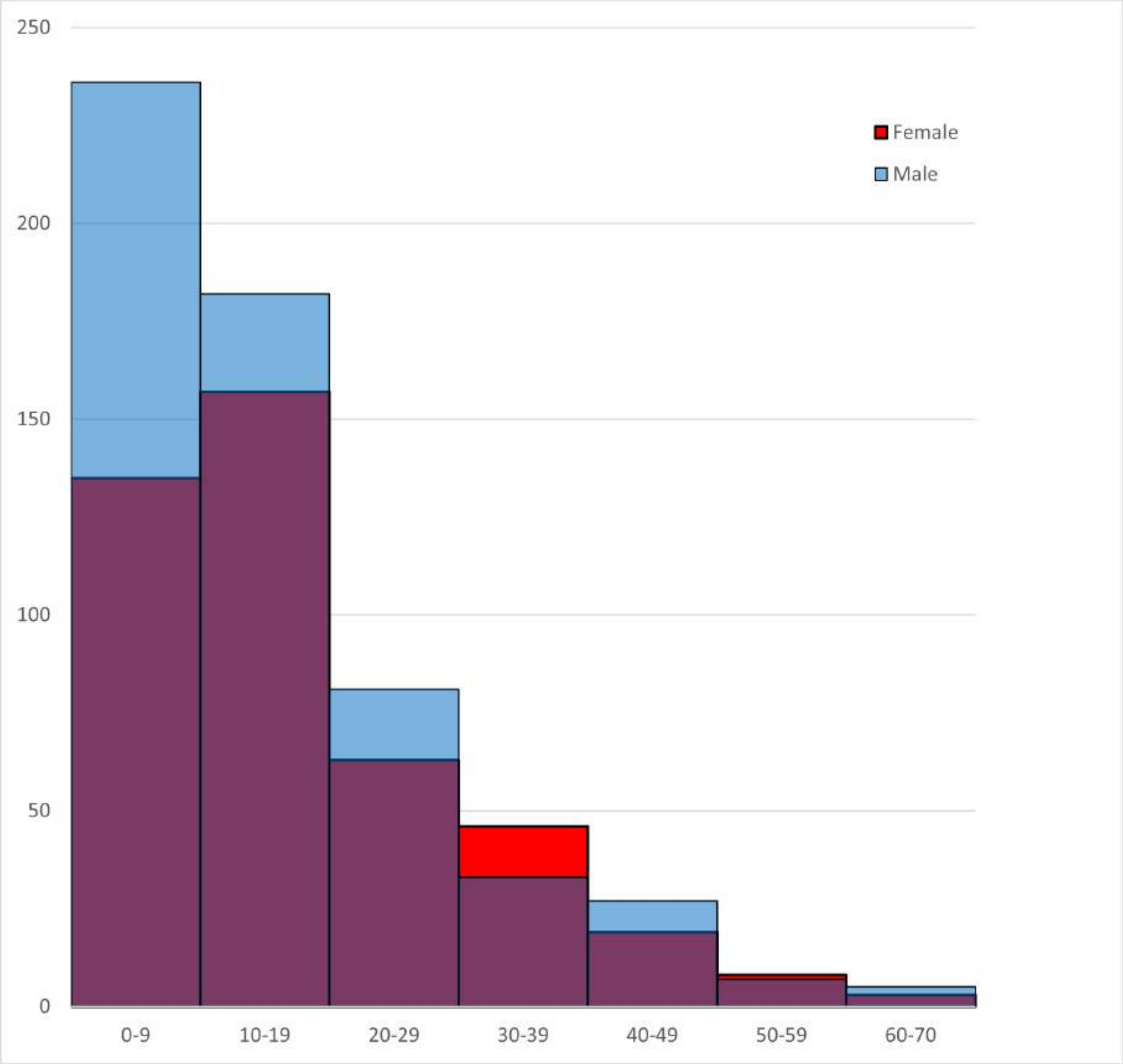


Figure 3.2: Distribution of Male and Female Subjects by Age - FG-NET. Red represents females, blue represents males, and purple is the overlap between both genders.

#### 3.3.2.2 MEDS I & II

Multiple encounter dataset (MEDS) [79] is a mugshot test corpus of 1309 male and female deceased subjects prepared by MITRE. MEDS has been provided to assist the FBI in facial recognition tasks, forensic comparison, training analysis, etc. There have been two different releases of the dataset supported by the National Institute of Standards and Technology (NIST). The distribution of age, gender and ethnicity can be seen depicted in Figure 3.3. The ages were grouped in ranges of 10 years. The available ethnicity categories are Black, Unknown, American Indian, White and Asian. The images belong to an age range of 17 to 69 year-old subjects. There is a predominance of males throughout the whole age range. The only predominance of female subjects is for the American Indian group from ages 37-46. It can be seen that there is a peek of over 350 images belonging to young Black males.

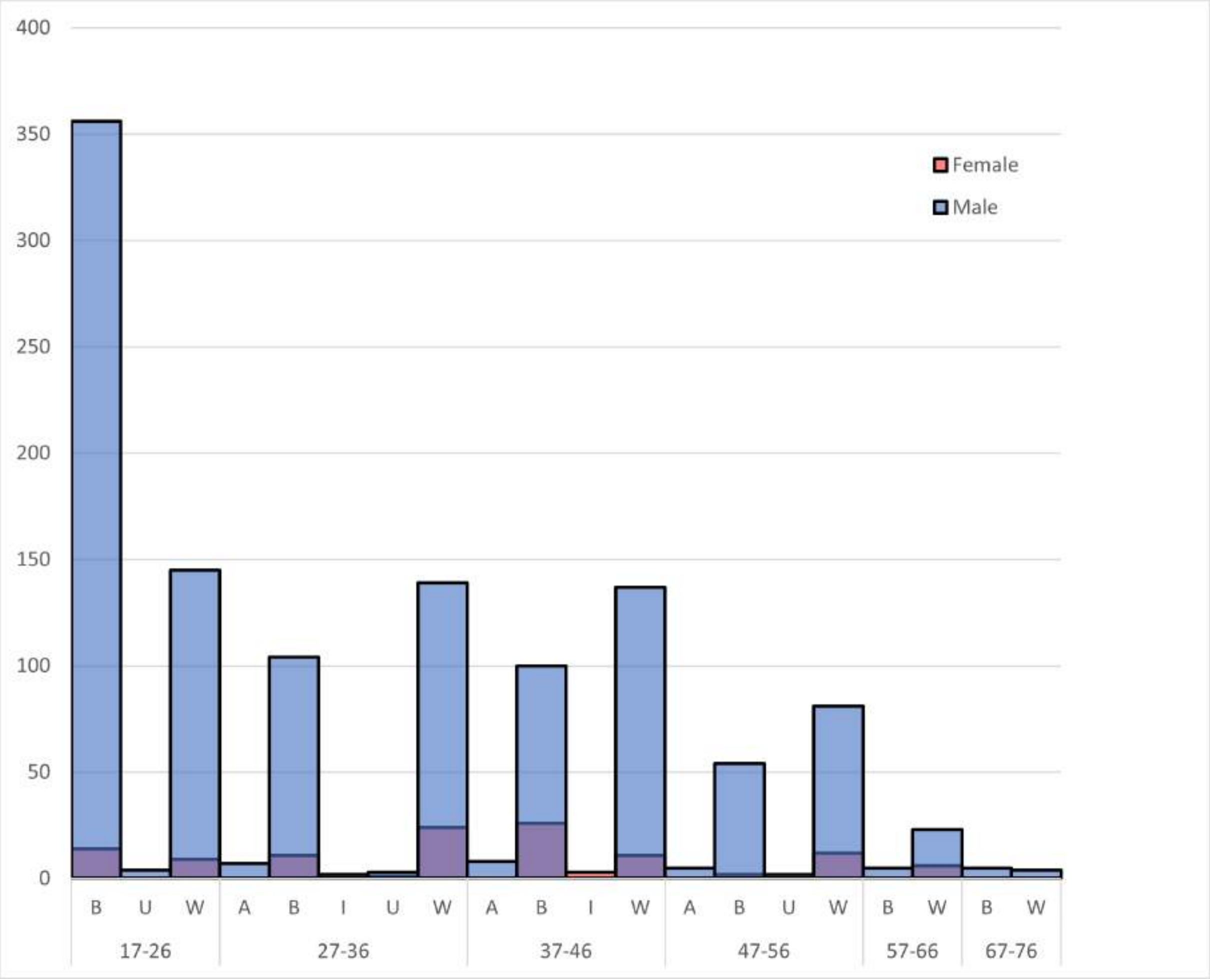


Figure 3.3: Distribution of Male and Female Subjects by Age and Ethnicity - MEDS. Red represents females, blue represents males, and purple is the overlap between both genders.

#### 3.3.2.3 FERET

The FERET dataset contains approximately 14,000 images and is pertinent to face detection [201]. The age was labelled based on estimates of the assessor. Our research cannot rely on conjectures due to the considerable MAE values for age prediction produced by the state-of-the-art age estimation algorithms. Therefore, this dataset was not used in our studies. Nevertheless, FERET has been widely employed in several facial recognition related studies. The ethnicity distribution can be seen depicted in Figure 3.4. A predominance of White subjects with a 62% of presence followed by an 18% of Asian subjects can be observed.

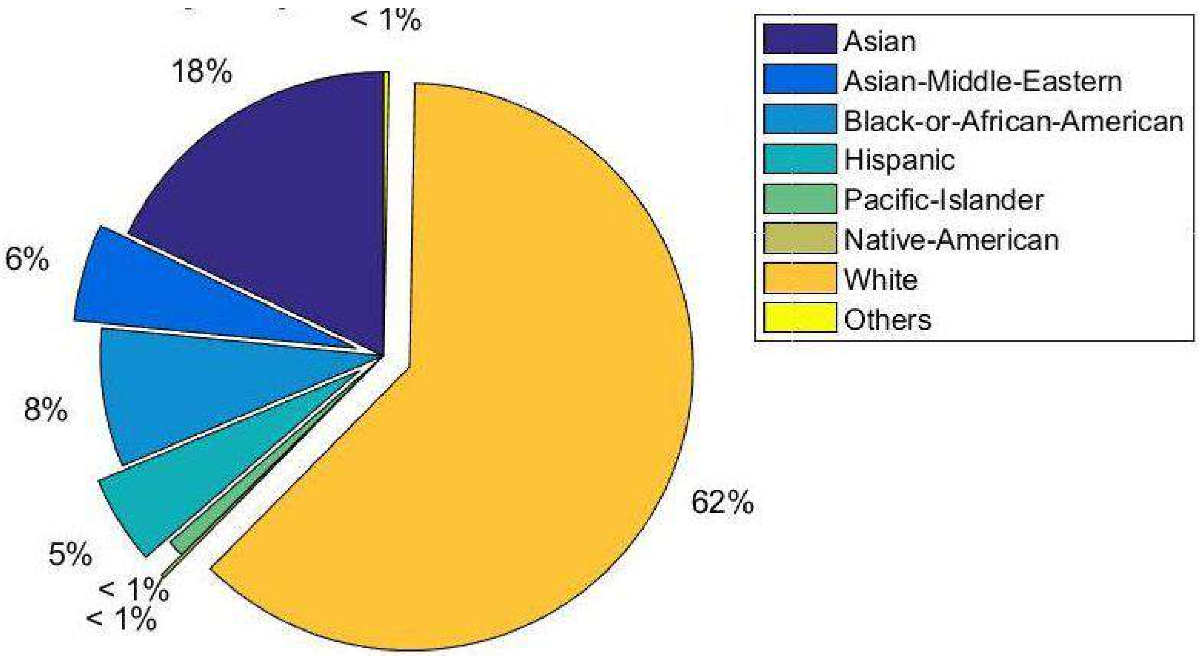


Figure 3.4: Ethnicity Distribution - Feret [178]

### 3.3.3 Medium Datasets

#### 3.3.3.1 Adience

The OUI-Adience set is a medium sized public collection of 19,487 labelled images obtained by online facial images of Flickr “in the wild”. The age labels were estimated by the assessors and are grouped in 8 classes: 0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+. Although Eidinger et al. [70] have stated that they use CC licenses for their images, we have detected from a sample of 10,842 images, that 89.55% are associated to images with copyright. Therefore, the use of such dataset has been avoided.

The distribution of male and female subjects by age range can be seen depicted in

Figure 3.5. The figure shows that there is a predominance of male and female subjects for the 25-32 age range class.

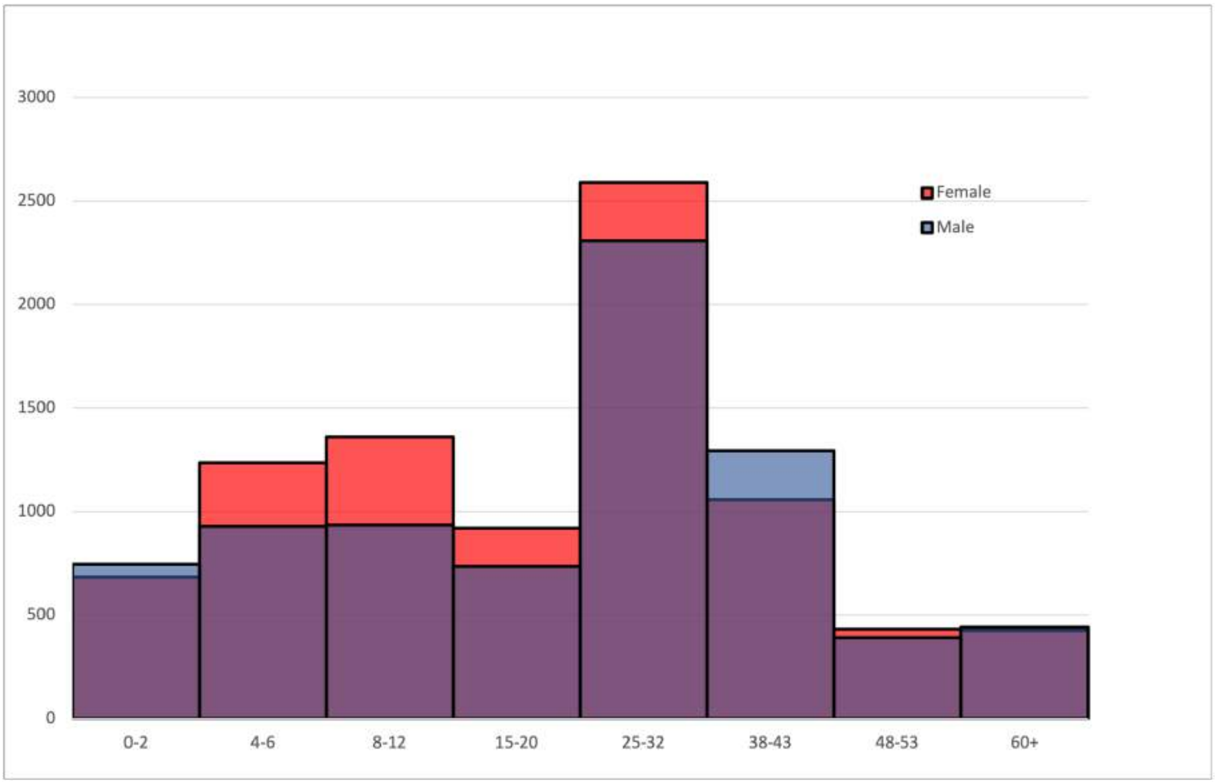
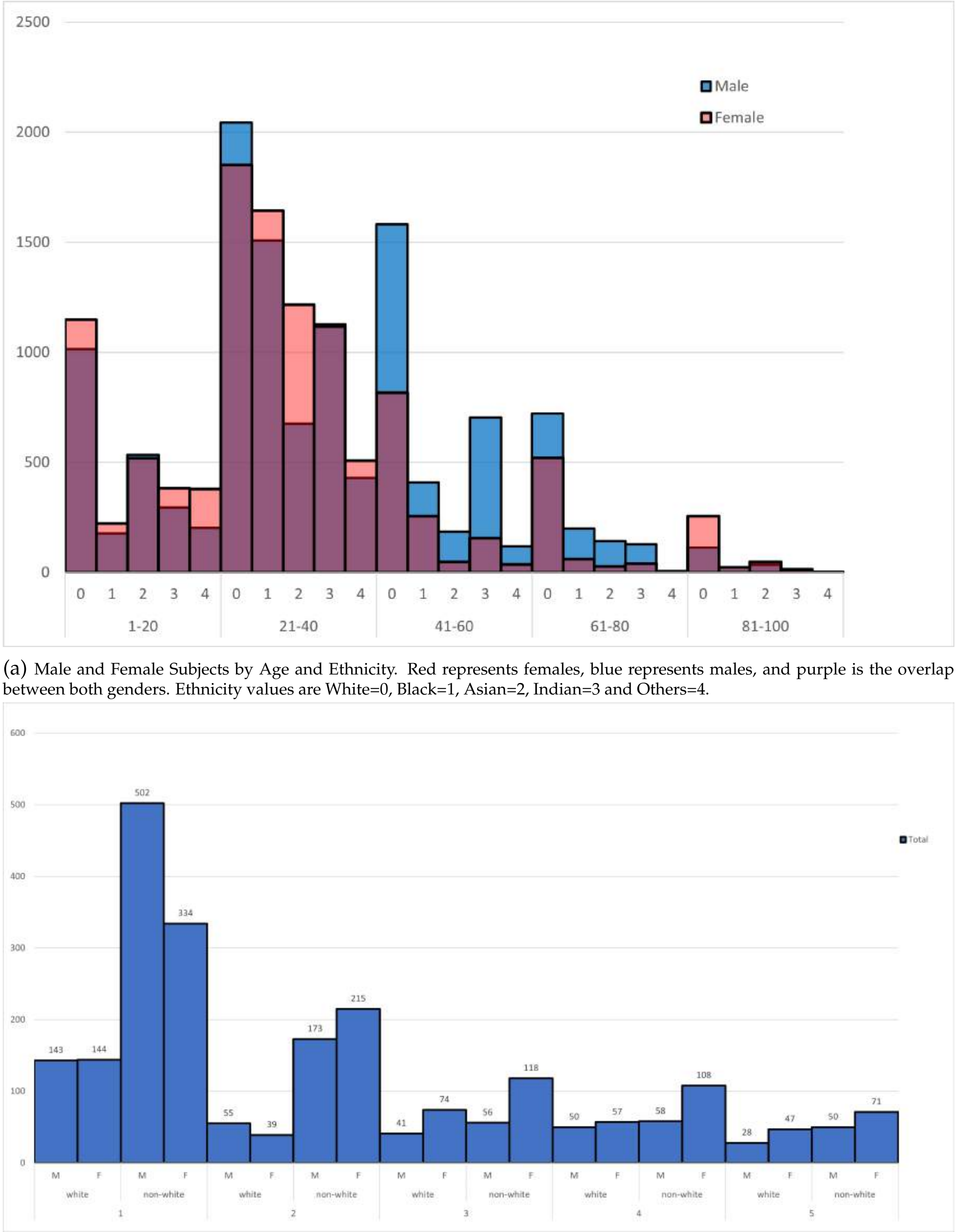


Figure 3.5: Distribution of Male and Female Subjects by Age Range - Adience. Red represents females, blue represents males, and purple is th overlab between both genders.

#### 3.3.3.2 UTKFace

UTKFace is a medium sized public collection that consists of over 20k images in a long age span (range from 0 to 116 years old), with age labels that have been automatically predicted with the offline DEX model. The images cover large variation in pose, facial expression, illumination, occlusion, resolution, etc. This dataset could be used on a variety of tasks: face detection, age estimation, age progression/regression, landmark localisation, etc. [270]. The age, gender and ethnicity distribution can be seen depicted in Figure 3.6a. The ages have been grouped in bins of 20, the gender is denoted by 0 = male, 1 = female. The ethnicity is an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern) respectively. It is observed that there are predominantly female subjects throughout the age ranges. The main concentration of subjects lies in the 21-40 age bin. For subjects less than and equal to 20, the main ethnicity present is White. In Figure 3.6b, the first five age classes can be seen with the count of subjects per age, gender and ethnicity. The ethnicity was grouped in White and Non-White. It can be seen that there is a possibility to create a small test dataset balanced by age, gender and race/ethnicity of 30 images per class.



(a) Male and Female Subjects by Age and Ethnicity. Red represents females, blue represents males, and purple is the overlap between both genders. Ethnicity values are White=0, Black=1, Asian=2, Indian=3 and Others=4.

(b) Subjects per Age, Ethnicity and Gender. M represents male and F represents female.

Figure 3.6: Demographic Distribution - UTKFace

### 3.3.4 Large Datsets

#### 3.3.4.1 Selfie-FV

Selfie-FV is a large dataset of facial vectors derived from unique face images of subjects between 8 and 38 years old. The size of the collection also exceeds 21k subjects and the data is shared in two pickle files: one for training and the other for testing (no-one appears in both the training and test sets). The files were created with a 80/20 split and contain the pickled Pandas Dataframe objects with a unique identifier for the image, the image number, the image filename, the accurate age ground-truth and the facial embedding. The dataset was prepared by collecting selfies published by celebrities with known dates of birth [9] The files are available on Github: `https://github.com/EdwardDixon/selfie-fv/`.

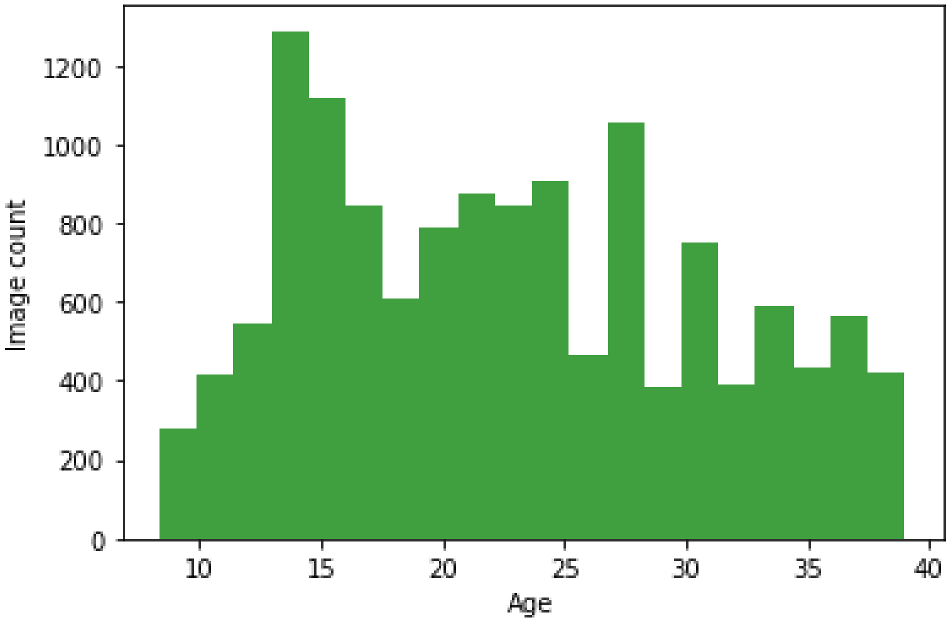


Figure 3.7: Distribution of Female Subjects by Age - Selfie-FV [63]

The age distribution can be seen depicted in Figure 3.7. It can be observed that while the amount of images is increasing from age 8 onwards, a sudden peak occurs close to the 15 year old mark. These images are in the age range of interest to develop an underage age estimation model.

[9] `https://www.famousbirthdays.com/`

#### 3.3.4.2 MORPH

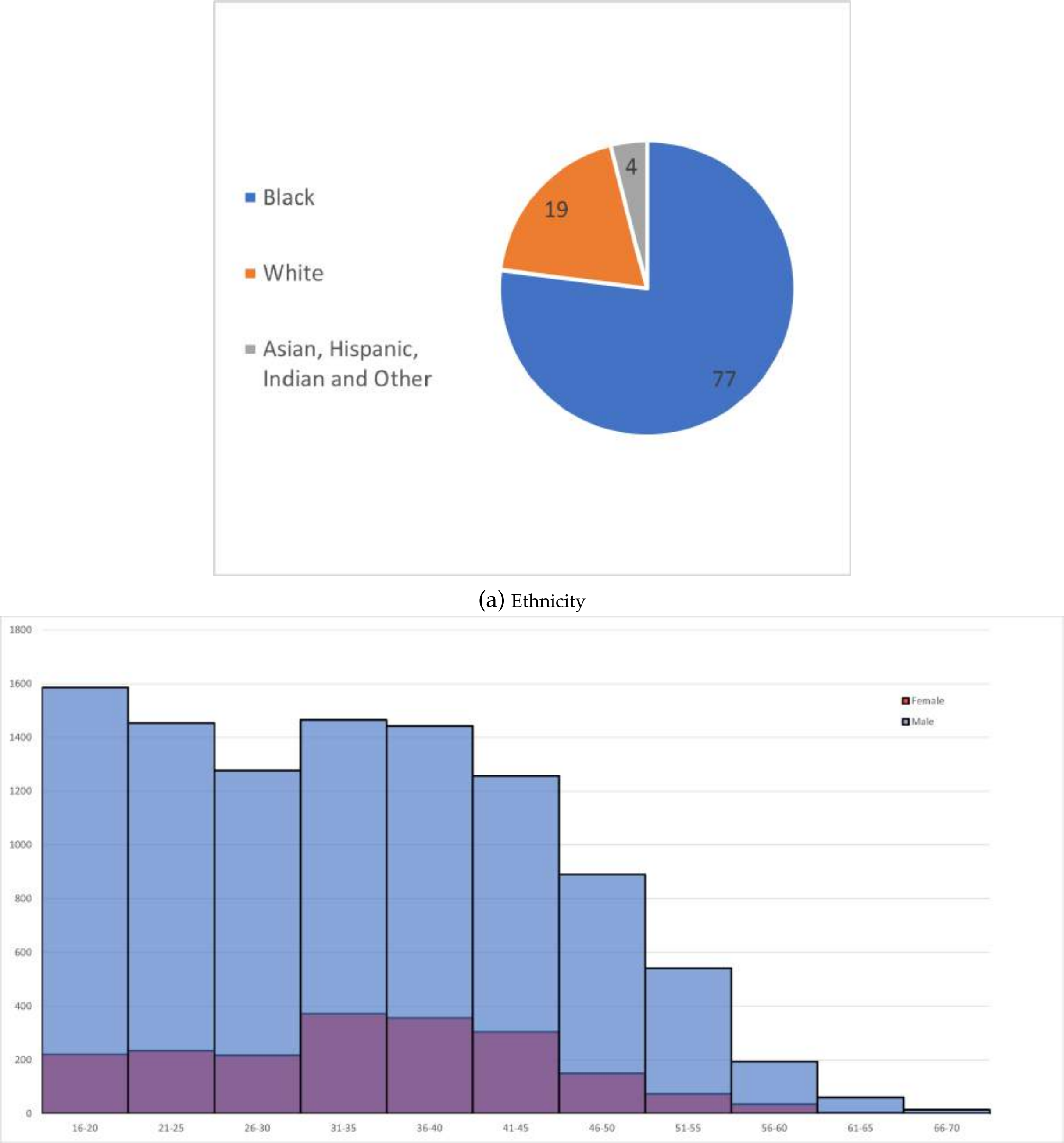
The MORPH dataset is categorized as a large dataset and consists of approximately 78,000 images of subjects but only 55,134 images are non-commercial and available for academic purpose. The age ranges are between 15 and 77. This dataset is useful for facial recognition and relevant to our work when we consider age estimation for the teenager age group. Age and gender labels are accurately documented. The images are mugshots that contain age metadata accurate to the year. The ethnicity distribution can be seen depicted in Figure 3.8a. The faces corresponding to Black ethnicity account to 77% of the dataset, while the White faces equate to 19%. The rest (4%) correspond to Asian, Hispanic, Indian and Others.

In Figure 3.8b the predominance of male subjects can be seen depicted throughout the entire age range. The most concentration of images are in the age range 16 to 45. After 45, the age counts plummet exponentially. For the age range close to the borderline between childhood and adulthood (16 to 20), the number of subjects is the highest within the dataset. This age range is important for forensic investigations.

The size in pixels of each image is 400 x 500. The distribution of image quality can be seen depicted in Table 3.1. The image quality field was designed by Ricanek and Tesafaye to give researchers the option of choosing images of a certain degree of quality. Although image quality can be highly subjective, the authors employed an unbiased rating system [212].

| | **Poor** | **Fair** | **Good** |
|---|---|---|---|
| **All** | 8.40% | 48.70% | 42.90% |
| **Males** | 6.70% | 39.70% | 35.20% |
| **Females** | 1.80% | 9% | 7.70% |

Table 3.1: Distribution of Image Quality - MORPH [212]



(a) Ethnicity

(b) Age and Gender. Red represents females, blue represents males, and purple is the overlap between both genders. In this graph, female subject counts never surpass males. Hence only the overlap can be seen.

Figure 3.8: Demographic Distribution - MORPH

#### 3.3.4.3 IMDB-WIKI

IMDB-WIKI is the largest public face dataset with computed age and gender annotations [218]. It has been subject of hundreds of facial recognition studies. The images were scraped from thousands of celebrities in IMDB[10] and correlated with Wikipedia[11]. The collection is quite considerable as the figures reach over half a million with an ample age range. Nevertheless, the calculation of age is acknowledged by the authors to not be entirely accurate. We have corroborated that there are inaccurate age labelling and presence of noise. Furthermore, we have taken extra care in using these images due to copyright restrictions. In Figure 3.9 the age distribution can be seen depicted. There are a lack of images in the underage age group. The images reach a peak for subjects around 25 years old and then the images gradually decrease.

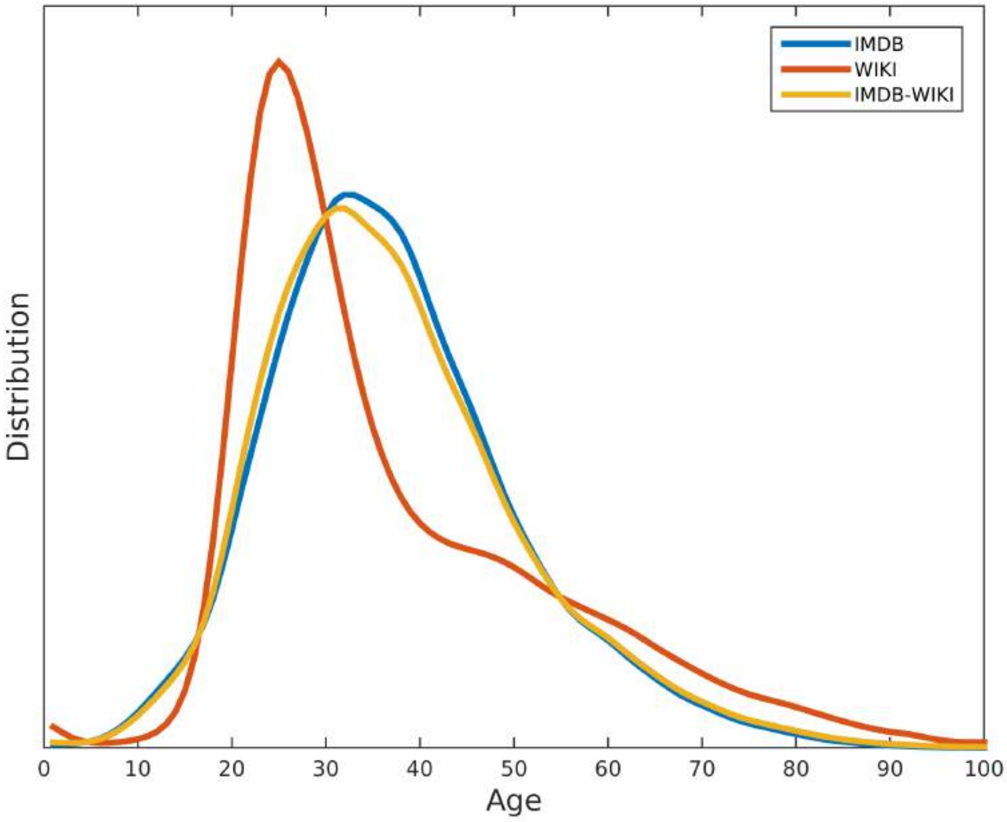


Figure 3.9: Distribution of Age - IMDB-WIKI [218]

[10] https://www.imdb.com/

[11] https://www.wikipedia.org/

#### 3.3.4.4 Yahoo Flickr Creative Commons 100M

Another dataset that uses Flickr as a source is the Yahoo Flickr Creative Commons 100M (YFCC100M) that was released in 2014 [239]. This is the biggest dataset of images and videos publicly available for researchers. Due to the size of the collection and the dataset being distributed solely as the metadata, the database is constantly evolving, i.e., the photographs need to be downloaded individually from Flickr. The age and gender distribution can be seen depicted in Figure 3.10. The creators of the dataset annotated the age and gender labels using the automated DEX model. It can be observed that there is a scarcity of images in the age range of 0 to 3. The predominant group of images are present in the age range 20 to 30. Regarding gender, the male group is prevailing.

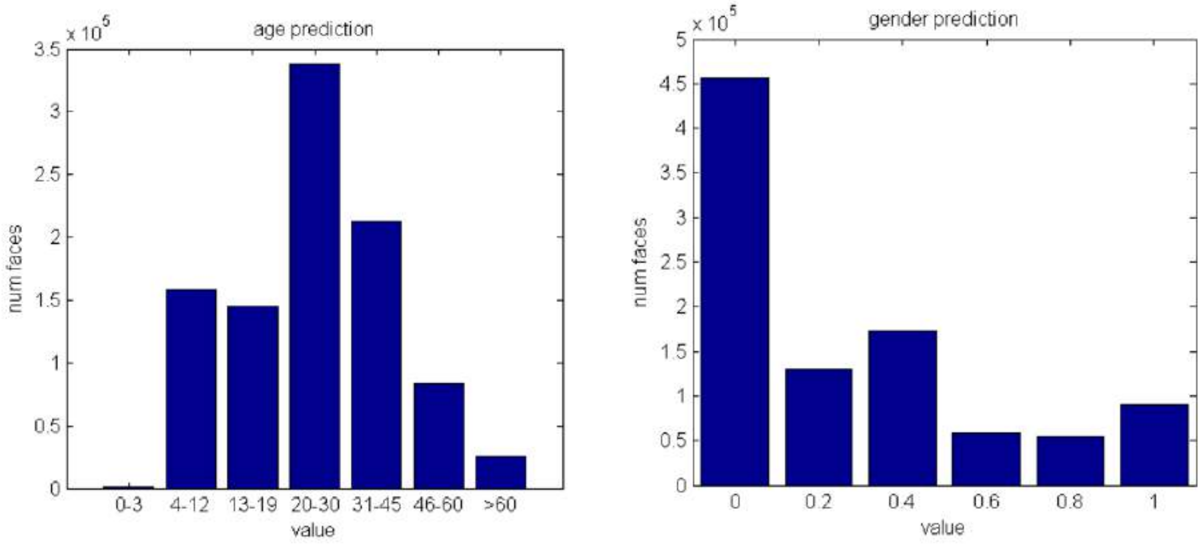


Figure 3.10: Distribution of Age and Gender - YFCC100M [174]

### 3.3.5 Underage Facial Age Datasets

Large datasets for underage subjects with accurate labels are rare; furthermore, the amount of children facial datasets available to the academic community is limited. Accurately labelled age and gender datasets are preferred over apparent age estimation. The only dataset that does not consider children and was discussed in this dissertation, is the MEDS dataset.

In 2013, Dalrymple et al. created a set of images with the following variations: 8 facial expressions, 5 angles and 2 lightning conditions. The collection consisted of combinations of these variations for 40 male and 40 female Caucasian children between 6 and 16 years-old. The real age was documented and also estimated by external raters with a 79.7% accuracy. Later in 2014, a 50 image dataset of female subjects aged 10 to 19 years from Germany, Italy, and Lithuania was created. In 2015, the In-The-Wild Child Celebrity (ITWCC) dataset was created by Ricanek et al. and the set was composed of 304 subjects with a total of 1,715 images (876 female and 839 male) from the age range 5 months to 32 years. Next in 2016, the Boys2Men collection was created as a private dataset that mainly focused on male images from the age range 12 to 21 years-old [41]. In the same year, ageCFBP (an underage-focused face dataset) was created. The parents/guardians of the subjects gave consent for the use of their children data. [96]. Following in 2018, Deb et al. released a dataset containing 3,682 face images of 919 subjects, in the age group 2 to 18. Each subject has an average of four images that have been acquired over a time span of approximately 4 years. The dataset is comprised of 66% boys and 34% girls [56]. It is notable that in the past 5 years, the number of underage datasets has grown but still requires validation, accurate age labels, and balance.

### 3.3.6 Data Pre-Processing Stage

An important step for an image classification task is to filter unnecessary features that would affect the learning process of a ML algorithm. In some cases, facial image preprocessing may not be necessary if the source is akin to a standard passport photograph. However, facial images in-the-wild may have characteristics such as various pitch/roll/yaw angles, multiple subjects per image, background noise, varying image size and quality, etc. Such photos require image pre-processing and normalisation to align and remove unnecessary features. Reisfeld and Yeshurun suggest that the knowledge of the location and the scale of a face impacts positively on the speed and reliability of face recognition systems.

In 2015, Han et al. [102] designed a face pre-processing procedure to overcome image variations due to external factors. Their approach entails: (1) converting a colour facial image into greyscale, (2) rectifying the face based on the two eyes and cropping to 60x60 pixels with a 32-pixel interpupillary distance (IPD), (3) detecting the face and the eyes using Cognitec's commercial FaceVACS SDK, and (4) applying the Difference of Gaussians filtering. In the same year, Liu et al. [159] proposed a deeply learned regressor and classifier for robust apparent age estimation. A three step preprocessing procedure was implemented: face detection, facial landmark localisation, and facial normalisation. For the first, a face detection toolkit developed by VIPL lab of CAS was used; for the second, 5 facial landmarks were detected with a Coarse-to-Fine Auto-Encoder Network. For the last step, external and internal normalisation approaches were considered.

## 3.4 Data Bias and Ethical Controversy

One common application of data-driven methods is facial recognition. Several researchers have used large datasets to train models that can verify and identifying an individual from a set of images [199]. Many of these employ supervised learning and precise image property labels. However, it has been continuously reported that datasets can be biased. For example due to backgrounds and hair style [134]. A recent study in 2019 shows that most existing large-scaled face databases are biased towards lighter-skin faces compared to darker faces and thus lack diversity [174]. It has also been reported that datasets predominately contain weak label accuracy [267]. Public facial datasets are strongly biased towards Caucasian faces and other races despite the sheer amount of available data [131].

In the 1990s, Carter and Nixon [39] realised the need for large, well-documented, high-quality databases of humans, particularly for research purposes. However, Wang et al. [256] stated that due to the impracticality of collecting evenly distributed, wide racial/ethnic diversity of data, biased databases are more commonplace. Therefore, trained models and automatic human age estimation are unable to handle race/ethnicity and gender without bias and thus cause the performance to decline.

The influence of race and gender seems to be the most common as both of these attributes play an important role in age estimation. The Wang et al. [256] study offers a solution to minimise the influence of these factors in automated facial age estimation. It should be considered that models trained with unbalanced datasets will produce biased results thus leading to compromised accuracy and raising ethical concerns about fairness of automatic systems. This topic of study has emerged critically in the recent ML and AI literature.

Per Birhane [25], for any individual, community or situation, algorithmic classifications can be either advantageous or detrimental. It has been discussed that the use of facial age estimation can aid digital forensic investigations and be used in several other venues. But there is little to no research on the disadvantages of the use of such technology. Similar to facial recognition, age estimation may have an influence on systems that use facial age estimation to produce a social output and as suggested by Birhane, these AI systems may have an impact on vulnerable communities. For instance, it would be harmful that a model would fail to categorise appropriately, the age of an undocumented child asylum seeker belonging to a non-white ethnic background due to racial data bias.

Controversy has surrounded the use of AI for facial recognition especially as many of these systems have been shown to be biased in terms of race/ethnicity and/or gender.

For instance, facial recognition systems (embedded in most smartphones) perform better for those who are white and male [35].

With the use of facial recognition systems there are “fears of an Orwellian invasion of privacy” [32]. Implying surveillance issues, privacy concerns and lack of consent. This can be seen as a negative impact of the use of such technology, being declared in the UK as a technology that violates human rights, data protection and equality laws by the Court of Appeal [13] and deemed unlawful in several other states.

Facial age estimation has also shown bias towards ethnicity as discussed in Section 3.2.11.5 and the unfair practices that have been previously accomplished such as detaching a whole ethnicity to eliminate bias. The current main disadvantage of using age estimation is the accuracy for underage subjects regardless of race/ethnicity and gender.

## 3.5 Summary of Literature Review

There are hundreds of studies of facial age estimation but only the high performing approaches that report MAE as the model metric are presented. This way it is straightforward to perform comparisons. Only offline age estimation models were presented due to online models being termed as “black boxes”. Hence, not being possible to appreciate why certain online model had a better performance than another one. Facial datasets relevant to age estimation were compiled. Each dataset was downloaded and certain demographics were obtained. The best performing facial age estimation approaches per year and the facial datasets for age estimation are summarised in Tables 3.2 and 3.3.

| year | alg. | dataset | MAE | source |
|---|---|---|---|---|
| 2002 | R | Private | 4.3 | Lanitis et al. [147] |
| 2004 | R | Private | 3.82 | Lanitis et al. [148] |
| 2005 | R | FG-NET | 5.81 | Zhou et al. [272] |
| 2006 | R | FG-NET | 6.77 | Geng et al. [88] |
| 2007 | R | FG-NET | 5.33 | Yan et al. [264] |
| 2008 | R | YGA | 4.38 | Yan et al. [265] |
| 2009 | C | FG-NET | 2.58 | Guo et al. [98] |
| 2011 | H | FG-NET | 4.1 | Luu et al. [165] |
| 2012 | H | FG-NET | 5.9 | Wu et al. [262] |
| 2013 | R | MORPH | 4.0 | Guo and Mu [100] |
| 2014 | C | FG-NET | 4.5 | Han and Jain [103] |
| 2015 | C | FG-NET | 1.3 | Li et al. [154] |
| 2016 | C | MORPH | 2.78 | Hu et al. [112] |
| 2017 | C | MORPH | 2.56 | Rodríguez et al. [216] |

Table 3.2: Best Performing Offline Approach per Year. Algorithm Types are Regression, Classification and Hybrid.

| Dataset | Size Cat. | Size | Age Range | Face Type | Image Quality/ Size avg. | Distr. |
|---|---|---|---|---|---|---|
| FG-NET | small | 1002 | 0-69 | Frontal | 405 x 497 | Fig. 3.2 |
| MEDS I/II | small | 1309 | 17-69 | Frontal/ (Left/ Right)/ (profile/ angle) | 498 x 605 | Fig. 3.3 |
| **FERET** | small | 14K | 10-60+ | Frontal | 24-bit color | Fig. 3.4 |
| **Adience** | medium | 19K | 0-60+ | LFW | Smart-phone | Fig. 3.5 |
| **UTKFace** | medium | 20K | 0-116 | Aligned/ Crop | 200 x 200 | Fig. 3.6a |
| Selfie-FV | large | 21K | 10-38 | Selfies | Face Vectors | Fig. 3.7 |
| Morph | large | 78K | 15-77 | Frontal | Tab. 3.1 | Fig. 3.8a |
| IMDB-WIKI | large | 500K | 0-100+ | LFW | 500 x 500 | Fig. 3.9 |
| ***YFCC100M** | large | 1M | 0-60+ | LFW | 256 x 256 | Fig. 3.10 |

Table 3.3: Summary of Facial Datasets for Age Estimation. The Face type is categorised as frontal, left and right profile, left and right pose, and labelled from wild (LFW). The Distribution column has the references to the distributions pertaining to each dataset. *YFCC100M has 100M images but only around 1M are single face. Age has been predicted for datasets that are in **bold**.

CHAPTER

FOUR

# METHODOLOGY

The methodology to address the research question "how can age estimation be improved for digital forensic investigations?", is described in this section. This question leads to the development of several software components to assist the data curation, model implementation and evaluation. Hence a building methodology was applied. To be considered research, the construction of the artifacts must be new or include new features that have not been built before. The dataset generator is a novel approach that creates balanced datasets out of the input of several existent facial datasets. The VisAGe dataset is presented as the largest curated underage age/gender labelled facial dataset. It is the outcome of an automated voting software process where the age and gender labels are assigned by consensus among human annotators. This dataset was evaluated and a new facial pre-processing technique was developed. In recent years, AI based approaches for automated age estimation have been created, and many public cloud service providers offer this service on their platforms. The accuracy of these algorithms have been improving over time. These existing approaches perform satisfactorily for adult subjects, but perform wholly inadequately for underage subjects. An experimental methodology was used to evaluate new solutions for the underage age estimation problem. Several models were developed and in the making, a facial pre-processing technique based on an artistic facial proportion approach was conceived. Finally, two different approaches were employed to tackle the underage age estimation problem: classification and regression. Both models were experimental methodologies are linked to the same research question previously mentioned and require performance measurements.

From the beginning of this research, there was an interest to include ML and DL as tools and techniques to gather and process evidence in a timely manner. Inspired by the evergrowing backlog that has increased throughout the years and also, the desire to assist the most vulnerable groups, improving the accuracy of age estimation was a major challenge; due to the nature of courtroom practice, and the necessity of expert testimony, it is neither intended nor anticipated that these AI techniques will fully replace trained investigators. Rather, this type of investigative aid has the potential to greatly expedite digital forensic analysts in their work, and potentially lower the psychological load of dealing with CSEM material on an ongoing basis.

The first step to improve age estimation was to survey the existing tools/techniques and establish a baseline analysis of performance evaluation. The establishment is discussed in Section 4.1. An overview of the performance of several offline age estimation models are presented. Cloud-based age estimation services are compiled and exhibited, and the alternatives to access cloud-based resources are shown. This sections further leads to the detection of low performance in certain age ranges, specifically for the underage age group. A need for balanced datasets with low performing age range was noticed and therefore, an initial prototype to gather balance images from several datasets such as the ones discussed in Section 3.3 was designed. To address performance issues, an early approach was developed and the implementation of ensemble learning was achieved. Through the timeline of this dissertation, the early approach was improved with several methods which are discussed in this section.

Finally, the assessment of the influencing factors of age estimation is discussed in Section 3.2.11.

## 4.1 Establishing a Baseline for Facial Age Estimation Accuracy

The initial step to take into account is the state-of-the-art facial age estimation models; due to the increasing applicability of DL in several fields, the advancements are also gaining progression on DF. It is paramount to monitor the evaluation model metrics (refer to Section 2.4.4) present in each research and the kind of dataset that has been used. Some researches may present remarkable performance but the model may not be able to generalise well due to the dataset being too small and biased. Other models are engendered from facial age images grouped by several ages. Usually tackled by a classification problem, the performance may be irrelevant for cases where the exact age is required. Moreover, the determination of the exact age while being affected by several external factors is demanding.

An introduction to facial recognition and facial age estimation has been covered in Chapter 3. A facial recognition task is related to an age estimation task due to the similarities used in the data although the former requires only faces and the latter is challenged by the scarcity of labels of age per face. Usually both tasks demand face detection (discussed in Section 3.1.3.2 and Section 3.1.3.1) so that the process can be completed faster; if a face is not detected, recognition is halted. The same would happen with facial age estimation.

There are two possible models for facial age estimation: offline and online which are discussed in Section 4.1.1 and 4.1.2 respectively. Following, open source models and cloud predictions services should be selected. To test how the model's performance improves over time a logging mechanism should be implemented or adapted. Due to the proliferation of big data and DL, a manual procedure is unfeasible. Instead, high-level programming languages may assist in the evaluation. Python supports several logging mechanisms from local libraries to cloud solutions such as neptune, tensorboard, etc. A distributed version-control system for tracking changes is necessary for these type of projects and a relevant relational or non-relational database is encouraged to store the metadata of the images from the several facial age datasets.

### 4.1.1 Offline Models as a Baseline

An initial point used for comparisons has been established with the revised literature, refer to Table 3.2 and Figure 4.1 for a summary of how the performance has evolved over the years in terms of MAE. The table contains the minimum MAE recorded per year of a group of studies performed over the same period. The trend throughout the years is decreasing at a steady rate and is predicted to be decreasing over time. Nevertheless, no matter how good the model is, the current exactness of age labelling would have an expected minimal error in the worst case of ±12 months. And due to age estimation challenges discussed previously, methods that achieve accuracy close to 100% are likely to be models that have memorised the inputs, have certain bias or the age classification problem has become too easy to solve, meaning that the generalisation to other datasets would not produce such great results, resulting in a poor quality model. Also in the aforementioned figure, it can be observed that there is a MAE fluctuation between 1.3 and 6.77. This range should be considered when evaluating newly developed models, or evaluating existing models with different datasets.

### 4.1.2 Online Services as Resources

Cloud-based facial age estimation is possible due to several companies allocating resources in face detection, recognition and age estimation. In particular the named "'Tech Giants': Microsoft, Amazon, IBM, Huawei, etc. In Section 3.1.3.2, online face detection was discussed and some of these same companies also provide attribute de-

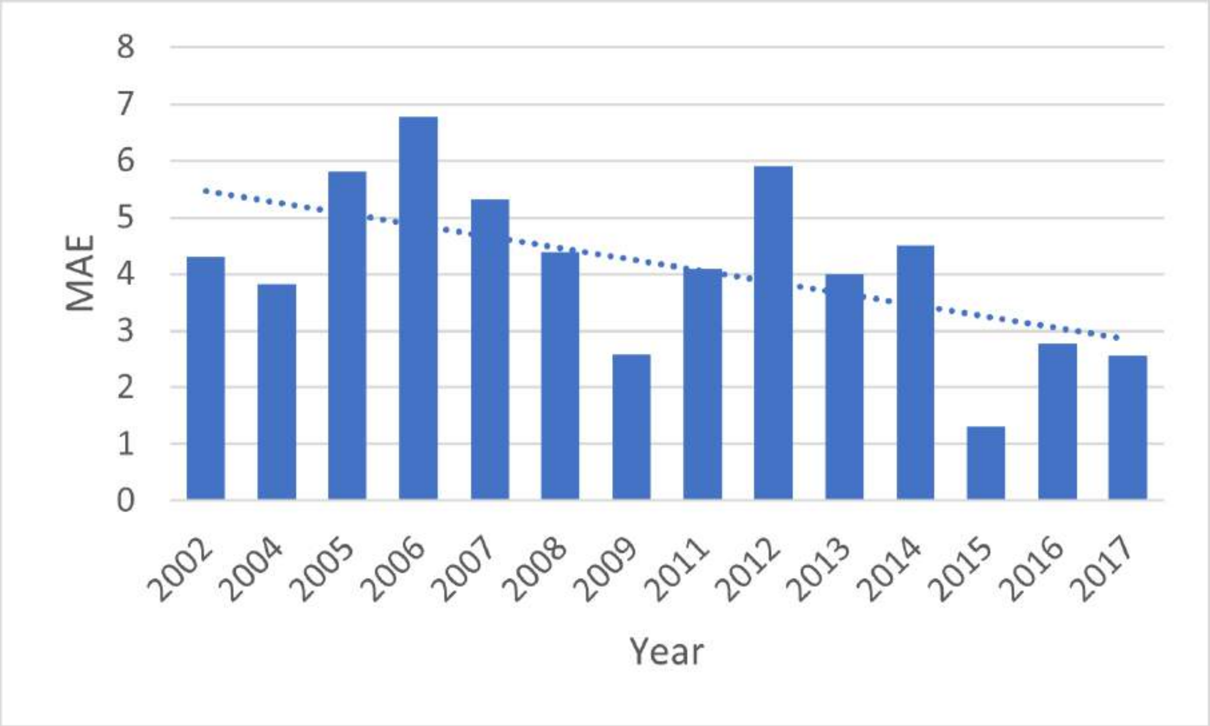


Figure 4.1: Minimum MAE per Year. The dashed line indicates a linear trendline that is decreasing at a steady rate.

tection which includes facial age estimation features but not all of them; i.e., Google does not provide currently age estimation services but did introduce an AI tool (freely available for non-governmental organisations and industry partners) to assist organisations in detecting and reporting CSEM online [8].

In Table 4.1, the most prominent cloud services that provide facial age estimation are listed. It can be seen depicted that there are several benefits from applying to a free tier plan. However, the number of images per month should be allocated conveniently to age estimation evaluation projects. For instance, several accounts to reach 5,000 images per month for the IBM Watson visual recognition API may be required. An alternative of the free tier would be the application for grants such as Microsoft Azure, Amazon Web Services (AWS) Research Credits, Google Cloud Platform (GCP) research credits program, etc.

[1] https://aws.amazon.com/rekognition

[2] https://azure.microsoft.com/services/cognitive-services/face/

[3] https://www.ibm.com/watson

[4] https://www.kairos.com

[5] https://www.faceplusplus.com

[6] https://www.how-old.net

[7] https://intl.huaweicloud.com

[8] https://www.betafaceapi.com

| Company | Service | Free Tier | |
|---|---|---|---|
| | | img/month | limitations |
| Amazon | Amazon Rekognition[1] | 5,000 | 1k metadata/month |
| Microsoft | Azure Face API[2] | 30,000 | $200 credit |
| IBM | Watson V. Recognition[3] | 1,000 | Deprecated |
| Kairos | Kairos[4] | n/a | Free 14 day tier |
| MEGVII | Face++[5] | ∞ | 1 query/sec |
| Microsoft | How-old.net[6] | ∞ | Image < 10 MB |
| Huawei | Customer A. Analysis[7] | n/a | 2k API calls are free |
| Betaface | BetafaceAPI[8] | 15,000 | 500 API calls/day |

Table 4.1: Cloud-Based Facial Age Estimation Services

### 4.1.3 Grant and Resource Access Timeline for Cloud-Based Instances

It is highly recommended to apply for several grant schemes that enable the use of facial age estimation due to the vast number of images required to be processed that at least 5,000 images per class should be gathered). Not all funds may be admitted but it is worth applying in a timely manner. As per Figure 4.2, the applications were granted in mid December 2017 until early April 2020, where access to the University College Dublin (UCD) Sonic Server was conceded. Access to the Sonic server is exclusive to UCD students and staff, and provides a cluster with limited restrictions. The server is equipped with 2 Nvidia Telsa V100's each and is convenient for processing high load of information.

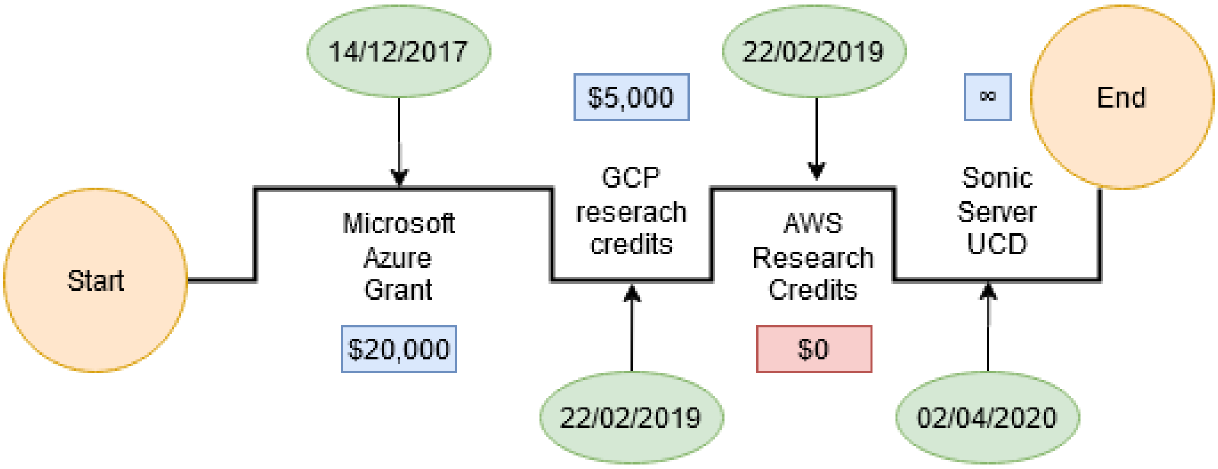


Figure 4.2: Cloud-Base Funding Timeline and Access to Resources.

### 4.1.4 Datasets with Age Labels for Bench-marking

Facial age datasets include IMDB-WIKI, FG-NET, MEDS, among others (for further reference, refer to Section 3.3). Combining and curating each of the aforementioned datasets together for age ranges 0-19 results in the total count of subjects per gender as depicted in Figure 4.3. It can be observed that per age the number of images are lacking in reference to the 5,000 images discussed previously. At the very most, 3,427 female images vs 2,516 male images for 19 year-old subjects are obtained.

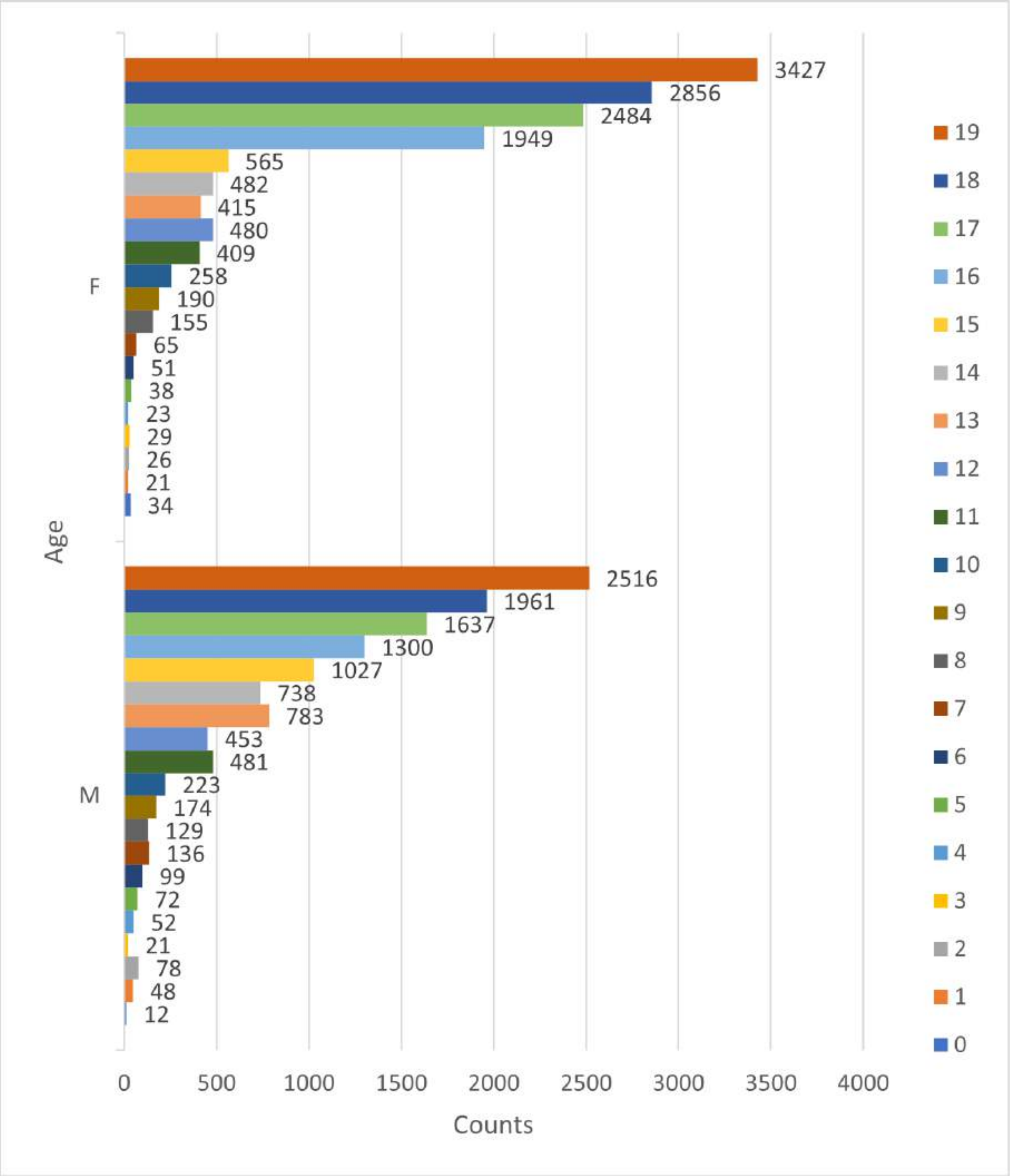


Figure 4.3: Curated images as a result of the combination of several facial age labelled datasets per age per gender.

## 4.2 Dataset Generation - Standalone

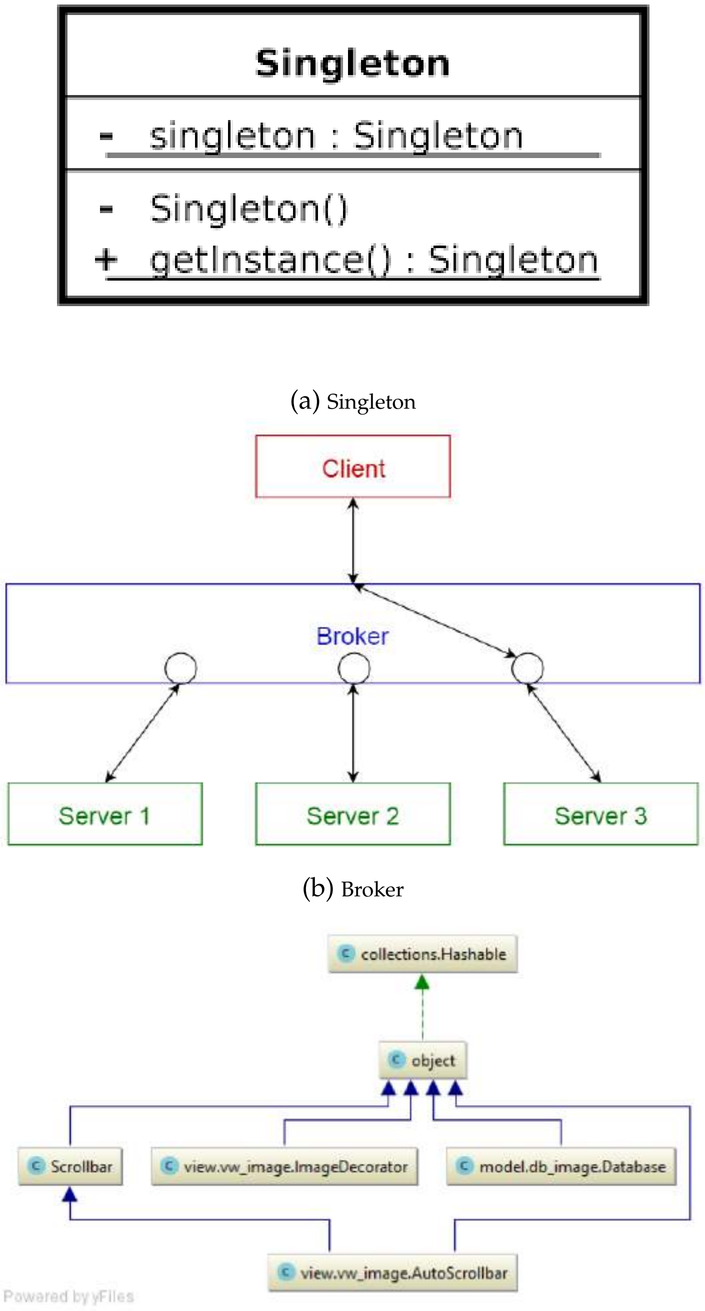
### 4.2.1 Overview of the Software Solution for Performance Evaluation

Evaluating the performance of several age estimation services requires a non-manual approach. There are several models that need the performance to be logged. An excel sheet would not suffice, the offline models require the integration with tools such as keras, caffe, tensorflow, etc. Online models require the consumption of an online service that would normally yield a JSON document as a response. This data requires to be stored in either relational or non-relational datasets. But there is an issue with datasets: they are either incomplete, biased, unbalanced or lack quality. To address this issue, the design of a dataset generator system was proposed. To evaluate the state-of-the-art cloud-based biometric services, a significantly robust set of labelled digital images was required. A non-biased collection of images needs to be generated by selecting random unique photos from the different datasets mentioned in Section 3.3. The query criteria applied to obtain the mentioned random images are: minimum age, maximum age, and number of images. The quality of the images is measured by subjective selection. The single user chooses the images that are deemed appropriate (frontal relevant images without noise. Refer to Figure 5.2). The user can also select the creation of a dataset of certain gender. But must consider that due to data bias, there could be lack of images for certain gender in certain age groups.

The software requires scalability to fulfil future adaptations to more datasets and services. Therefore, an model-view-controller (MVC)-architecture was proposed to enable code re-utilisation and parallel development.

### 4.2.2 Design Patterns

Multiple design patterns and inheritance are able to make scalability permissible. Inheritance enables the seamless addition of new services for future evaluations. Restrictions of instantiation are handled by the singleton software design pattern being useful when only one object is demanded to execute actions across the system. In this case, only one instance of a random set of images is necessary after the age criteria is applied. Refer to Figure 4.4a for the singleton unified modeling language (UML) class diagram. The decorator pattern allows adding additional features to an object without affecting the internal behaviour (Refer to Figure 4.4c). This is useful usually to implement additional functionalities related to the graphical user interface (GUI). The broker pattern may be used to structure distributed software systems with decoupled components that interact by remote service invocations. This design pattern can be seen depicted in Figure 4.4b. There is room for a plethora of design patterns; the ones mentioned are only a few out of a wide pool of decorator patterns available for software architecture.



(a) Singleton

(b) Broker

(c) Decorator

Figure 4.4: Design Pattern UML Class Diagrams

### 4.2.3 Image Database Design

A non-relational database is employed for simplicity of design and the image database contains all the dataset metadata stored there. The data structure per dataset varies but is quite similar in a sense that all the datasets contain a unique id set automatically by the database, each dataset manages the images with their own id with values such as *id, id_subject, subject_id, celeb_id and full_path*. The Wiki dataset uses the full path as an id. Another set of common attributes are age and gender. The only different dataset metadata for managing age is FERET, because it uses a year of birth (YOB) field but the age can be calculated with YOB and the *CaptureDate* attribute. Another important attribute is the image path: *URL, id, img_dname and img_name, full_path* are labels that are related to the image path. It is also observed that the *id* has been used in several occasions for the file path. Finally, a delete collection has a relationship with the dataset metadata, whenever an image from a dataset was discarded, it is recorded in the delete collection. All these descriptions can be seen depicted in Figure 4.5.

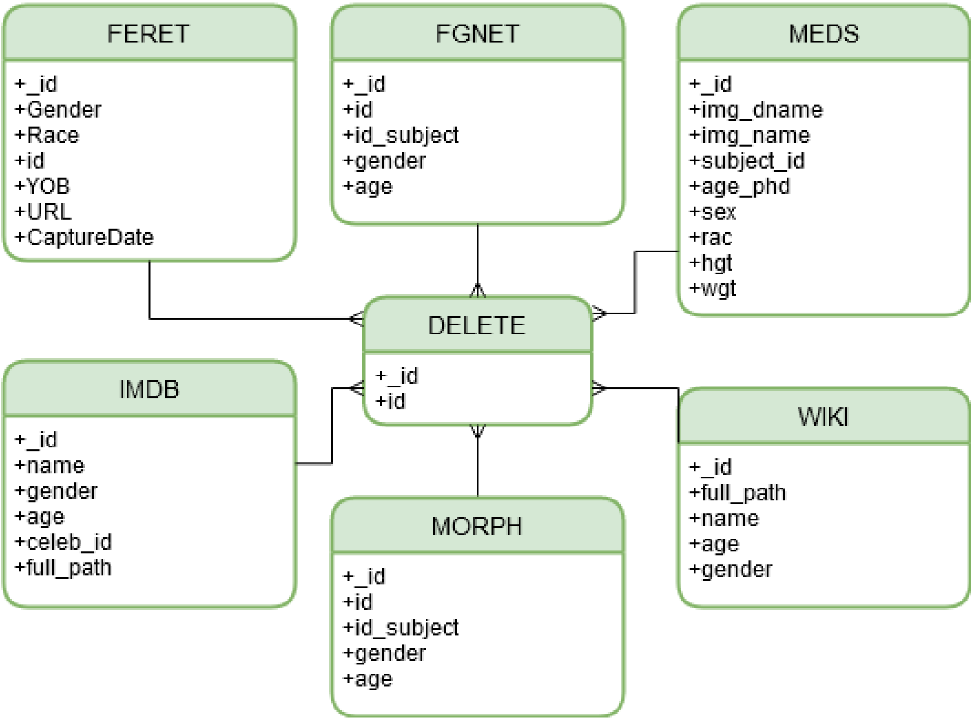


Figure 4.5: Database Design

### 4.2.4 Proposed Architecture

The proposed design of the system architecture can be seen depicted in Figure 4.6. The system was composed of multiple file sources such as the ones discussed in Section 3.3 (FG-NET, FERET, IMDB-WIKI, MEDS, YFCC100M, etc.). Each file source corresponded to an image dataset. The metadata was extracted from each image and stored in their corresponding collection. Scripts and Spreadsheets were used to inter-operate with non-relational repositories. Once a randomly selected dataset of images that is equally distributed by age and gender is engendered, both offline and cloud-based biometric services were automatically evaluated and the results stored in a common repository for further analysis. The final outcome was a graph in the form of a linear plot, box plot, bar plot, etc. It can be seen in the figure that the curation would be collection-based interacting with all the other components such as scripts, collections and data files.

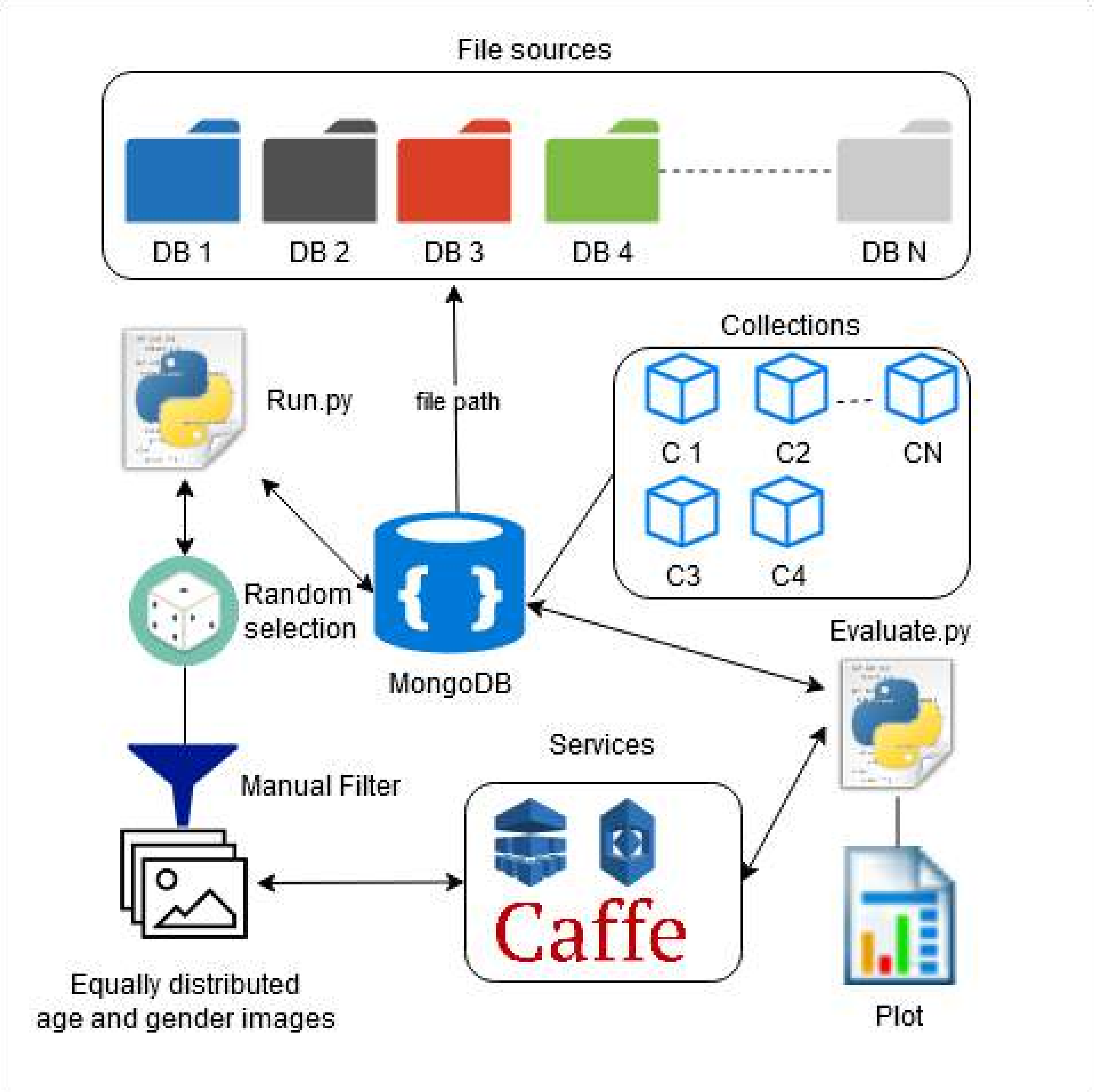


Figure 4.6: Dataset Generator System Architecture

### 4.2.5 Evaluation Performance Methodology

The type of evaluation used in the research was an empirical evaluation based on observation. The purpose of the evaluation is to find the least MAE within the different cloud-based biometric services and pre-trained models. The results of the evaluation would be helpful in selecting which one is most effective, or indicate what combination of different services might aid in creating a data fusion/ensemble approach. Our research exploited the pre-trained Caffe model produced by DEX to predict the age and gender of digital images. With this tool and other state-of-the-art online age predictors, comparative analyses were performed to evaluate the accuracy of several prediction platforms.

Each of the cloud-based biometric services mentioned for age estimation were evaluated by generating a random equally distributed dataset in order to avoid unbalanced analysis. An equal distribution requires the same number of images per class. This had a limiting effect on the total size of the dataset, as there was a lack of images for certain ages and as per Figure 4.3.

The goal number of images requested by the dataset generator may change due to a final quality assurance made by the user. There could be images missing in certain age ranges so the whole age classes would be modified in order to keep the dataset balanced; i.e., if a binary class consists of 50 and 45 images, to normalise the balance it would be easier and less time consuming to remove 5 images than to seek for 5 additional images. Nevertheless, when images are collected for academic purposes, they should preferably contain CC licenses with accurate age and gender labels.

Each image was evaluated by the offline and online age estimation services, and the results were recorded. Several influencing factors such as the error rates exhibited per entire age range, gender and age range groups, i.e. 0-9, 10-19, 20-29, etc. were scrutinised. The goal was to find whether certain systems perform better in different age ranges, or whether one system could be said to be the most accurate over the entire dataset.

### 4.2.6 Service Database Design

The data structure per service varies but is quite similar in a sense that all the services contain a unique id set automatically by the MongoDB management system, each service has its own JSON response and is stored in the collections. All the services are independent from each other but have a 1-to-many relationship with each dataset. The datasets used and the determination of the ages are described in Section 3.3. The service database design can be seen depicted in Figure 4.7.

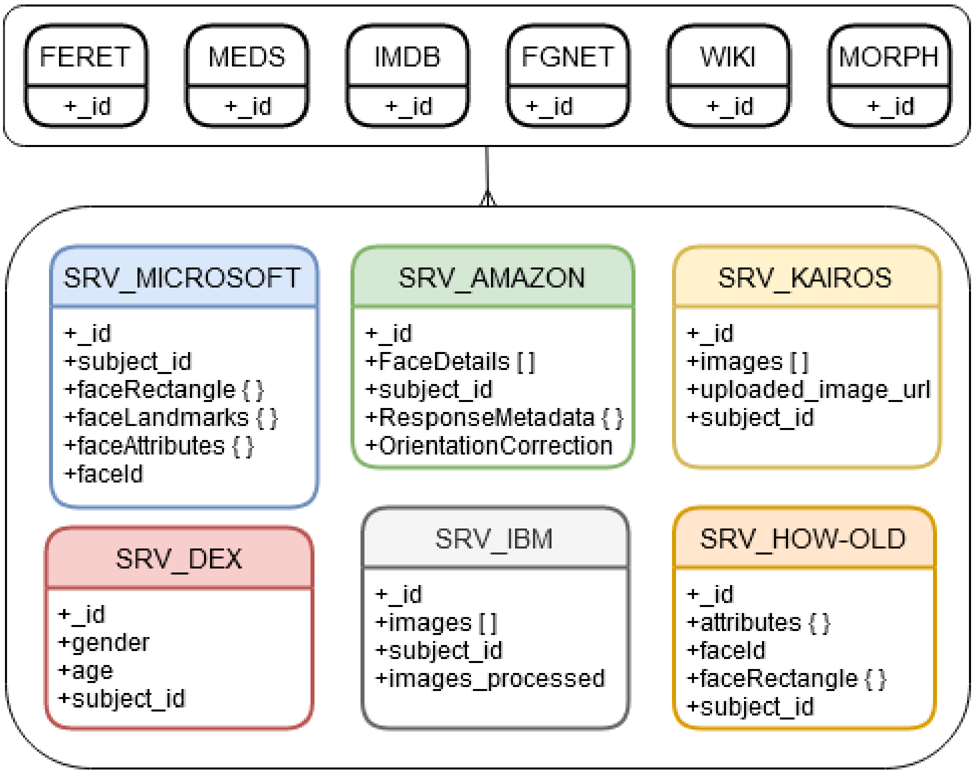


Figure 4.7: Service Database Design

Each service has a *subject_id* field which corresponds to the related record from a facial image dataset collection. It can be seen that each service has its own structure but they all have the age field which is of our interest.

## 4.3 Dataset Generation - Online

Specific trait labels such as age and gender are much less available in benchmark datasets, and they suffer from a lack of images in certain age ranges. The standalone dataset generator approach in Section 4.2 is able to curate images for a single user only. This not only being inefficient but time-consuming. In this section of our research we present VisAGe, a large-scale underage facial image dataset that merges human and machine annotations. Accurate age and gender recordings were the outcome of a voting procedure with several validations. The workflow of the data collection and curation is presented including the protocols used to assure reliability of the labelled data. The images collected can be used to train a more accurate age and gender estimation model for underage subjects, evaluate existing models and improve facial recognition.

## 4.3.1 Ethical Considerations

### 4.3.1.1 Access to Data

Online registration for access to the dataset is required due to the conditions of the ethical approval discussed in Section 1.2. The dataset is only available to established researchers in DF, cybersecurity, biometrics, forensic medicine, and related fields. The dataset and associated metadata is released under an attribution, non-redistribution license. The licence of the source photos remains unchanged from the original licence by the respective owners. Individual photo licences may/may not permit modification or derivative works, this is specified in the associated VisAGe metadata.

The access to the dataset requires registration that is available in the following link: `https://www.forensicsandsecurity.com/visage`

The VisAGe Dataset access request form is maintained by the data controller; the personal information requested is the institutional email, names, proof of existing experience/interest in biometrics, forensics, cybersecurity or related fields, and an explanation of the use of the dataset. A release agreement is provided and the requester must read and agree to the aforementioned agreement. The data controller will verify that the requesting person belongs to a respected research institution or research company. If the applicant is successful, a temporary link (single access; short 7 day expiration) to a data-source will be provided. While all photos contained within the dataset are licensed under public domain or CC licenses, and permit redistribution, individual photos may have restrictions regarding their modification or creation of derivative works. The license for each image (as taken from the source of the image) is included in the associated metadata. It is important to give attribution to the author when required, and to observe all conditions of the dataset license.

### 4.3.1.2 Creative Commons and GDPR

CC is a non-profit company that provides special licenses free of charge for the society that are alternatives for full copyright. The licenses enable the free distribution of material that would have been commonly subject to copyright. Flickr utilises 7 different types of licenses. For our research we collect images subject to 6 out of 7 licenses enumerated in the Table 4.2. It can be observed in this table that a copyright license can be used for research and scholarship for fair use. Nevertheless, copyrighted images are not used and the automatic generation of a file with attributions is created whenever necessary when using an attribution license.

| License | Copy | Distribute | Display | Perform | Condition |
|---|---|---|---|---|---|
| Copyright | N | N | N | Fair Use | Research/Scholarship |
| Attribution | Y | Y | Y | Copyrighted & derivatives | Give credit |
| Noncommercial | Y | Y | Y | Work and derivatives | Noncommercial |
| No Derivatives Works | Y | Y | Y | Verbatim Copies | No Derivatives |
| Share Alike | Y | Y | Y | Work and derivatives | Original license |
| Public Domain CC0 | Y | Y | Y | Work and derivatives | No restrictions |
| Public Domain Work | Y | Y | Y | Work and derivatives | No restrictions |

Table 4.2: CC Licenses used by Flickr

The Flickr users are liable for the images uploaded that have been obtained fairly and in accordance with the individual's rights, both key elements of GDPR compliance. In our research, our goal is to improve age estimation predictions. The data minimisation principle for our project is limited to the goal of reaching a MAE better than the current state-of-the-art techniques for the underage age group. For every image collected, we have the explicit consent of collection due to the CC licenses that has been configured by the owner of the photograph. Therefore, under article 6(a) which refers to "the data subject has given consent to the processing of his or her personal data for one or more specific purposes", our processing qualifies as lawful. Nevertheless, some safeguards have been implemented in accordance to the Human Research Ethics Committee of Sciences UCD. The storage device has been secured by encrypted disks, and personal information has been removed. Lastly, we have considered a protocol that will ensure the appropriate safeguards for data. This protocol includes the usage of a release agreement that has to been accorded with researchers that require our collected images.

#### 4.3.1.3 Anonymity & Disclosure

Previous to the entire data collection, a research ethical submission was issued. We have been granted the research ethical submission (LS-17-74-Anda-Scanlon) to use CC licenses to gather images of people from all age ranges. The respective approval document is attached on Appendix A.1.

Following the successful research ethical submission, the appropriate safeguards were applied to ensure the correct procedures for data protection. The station that collects the photos is secured by the principal investigator (PI) and only accessible within UCD through a secure shell (SSH) tunnel. The HDD is encrypted and any personal identifiable data such as geographical location was removed. To ensure the conditions for consent in the GDPR article 7.1 "Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data", a portable document format (PDF) is recorded with a copy of the

Flickr web page accessed the day the data was collected. Refer to the attached file on the Appendix B.1 for an extract of the PDF created with `wkhtmltopdf`[9].

### 4.3.2 VisAGe Design/Methodology

The VisAGe dataset collection system is presented in this section. For dataset collection, it is necessary to gather photographs of potential subjects, and subsequently use human judges to verify the age of these subjects. For this reason, it is necessary to ensure that it is possible to ascertain the correct age of each subject, rather than relying merely on human opinion. Consequently, the initial potential subjects are gathered by means of searching for images from birthday parties. Where the person whose party it is can be identified, this provides an accurate age. This also has the additional benefit of including subjects on (or close to) the beginning of each year of age, rather than at other points during each year.

The first step in the process is to identify and gather candidate images for inclusion in the dataset. Flickr is suggested as the source of images due to a number of factors. Firstly, it constitutes a huge repository of photographs with explicit license terms. In March 2019, Hidalgo reported that it hosted over 500 million public works licensed under a CC license [107]. Additionally, Flickr offers an API[10] that allows users to programmatically search for and download images and their metadata.

After searching for candidate photographs using the Flickr API, both the photo metadata and the images themselves were downloaded. In order to reduce the amount of human annotation required, an initial automated pass was conducted to attempt to identify the gender and other identifiable features of each subject, and filter images that do not contain a face or contain multiple faces.

Following this, an online judgement system was developed through which human judges voted on each image as to whether it was appropriate for inclusion in the dataset or not. This would become the methodology to validate the created dataset. After receiving three votes per image, those with three positive votes were included. For images with two positive and one negative vote, a second judgement process was undertaken to decide whether they are to be included. Images with two or more negative votes are immediately excluded from the dataset. Figure 4.8 presents the architecture of the system to facilitate this process. The structure of this system is divided in three main components, namely background scripts, data and web applications.

[9] `https://wkhtmltopdf.org/`

[10] `https://www.flickr.com/services/api/`

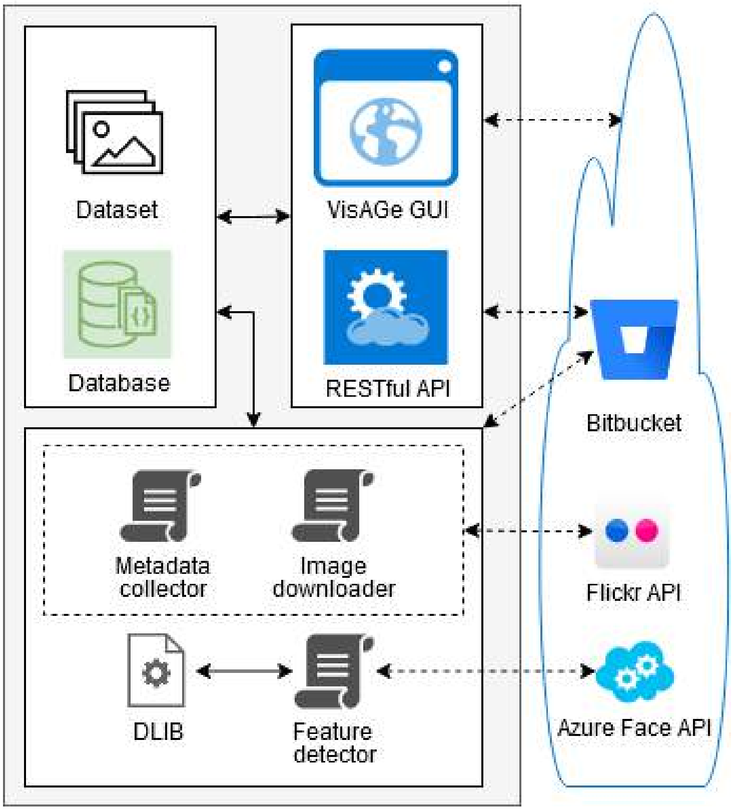


Figure 4.8: VisAGe Data Collection System Architecture.

#### 4.3.2.1 Photo Acquisition

The Metadata Collector and Image Downloader scripts interact with the Flickr API in order to gather the metadata and images of candidate photographs, based on a search.

The Metadata Collector connects to the Flickr API by sending valid credentials, a web method and the relevant arguments that allow a search query to retrieve data from an image. The search query can be crafted with the use of ordinals and the word *birthday*. For instance, a search query would send a "1st birthday" string to the Flickr search method and the outcome would be the image metadata of one-year-old subjects including a Uniform Resource Locator (URL) to the location of the original photograph. The image metadata may be stored in the database. This metadata represents data about the image provided either by the user or by Flickr itself, including user-supplied comments and tags about the photo and any album containing it, technical details of camera, resolution, date etc.

The image URL is then sent to the Image Downloader, which downloads the image, computes its MD5 hash and verifies this against the hash stored in the database. It also uses the MD5 hash to prevent duplicate downloads. Although Flickr offers several image sizes such as thumbnail, small, medium and large, the image quality that is collected is the size of the original photo as originally uploaded.

#### 4.3.2.2 Categorisation

Following the acquisition of the images and as part as the curation step, automated methods were used to filter and provide an initial categorisation. This was done via the facial feature detectors (as explained in Section 3.1.3), which employs `Dlib` and Microsoft Azure Face API [175] to obtain several facial features from the collected images. Azure provides the most complete annotations including facial attributes, landmarks, gender, age, etc. This is discussed in Section 4.6.1. Both face and gender detectors allow the classification of the images: the former discards multiple-face images and images not containing human subjects to retain only single-faced images and the latter segregates the photographs in either male or female folders. The gender classification is verified by the human judges at the next stage of the process, where they have the opportunity to correct them if necessary.

#### 4.3.2.3 Image Selection Criteria

The criteria to establish compliance considers several weights per variable and resembles the following formula 4.1.

$$\begin{split}((((age\_img * 0.6 + 0.2 * (age\_title|age\_tags)+ \\ subject\_bday * 0.2) * q) * no\_skin\_exp) * face\_single)\end{split} \quad (4.1)$$

Each variables and weights of formula 4.1 are explained in detail in Table 4.3. Both *quality* and *no_skin_exposure* are manually measured with a collective subjective methodology, and will be subject to automation in future work.

#### 4.3.2.4 Voter Background

The voters are males and females mainly from UCD and Mitre. The group is multicultural with backgrounds originating in Europe, Asia, North and South America. Diversity is key in the data labelling process.

| Variable | Weight | Description |
|---|---|---|
| age_img | 0.6 | In the image it is certain that the age is the true value. |
| age_title | 0.2 | The text mentions the age of the subject (implicitly or explicitly). E.g., Ted is 5 or Greg's first birthday. |
| age_tags | 0.2 | The text in one or more tags contains information about the birthday celebration subject, i.e., sweet sixteen. |
| subject_bday | 0.2 | It is certain that the image is the birthday boy/girl. |
| quality | 1 | The image is not blurry, and the quality is decent. |
| no_skin_exp | 1 | There is no considerable amount of skin exposure. Topless images are rejected. |
| single_face | 1 | Only single faces are allowed. An additional face is considered when at the very least, an eye, a nose and a mouth is visible. |

Table 4.3: Variables with the Corresponding Weights and Descriptions for Image Selection.

#### 4.3.2.5 Voting Procedure

A binary vote only allows a single answer from two possibilities. For instance, positive/negative. For a vote to be positive, a simple majority must be obtained (half the total number of votes, plus one). When the votes have been completed, if there is a disagreement but there are 2 positive votes, the image is analysed for a second time by an inspector.

As per Figure 4.9, there are three actors: voter, inspector and Q/A agent. The voting procedure is prone to human error or subject to disagreement; therefore, the voter has either made a mistake or the vote was qualified as disagreed. The inspector is the actor that reviews the disagreed images and then resolves the votes to either positive or negative. All the actors are able to delete images that are considered inappropriate for the dataset.

#### 4.3.2.6 Database Design

Since the majority of request responses are returned in a JSON format, it is convenient to store the information in a non-relational database. The database design can be seen depicted in Figure 4.10. The only independent entity would be the "release" collection

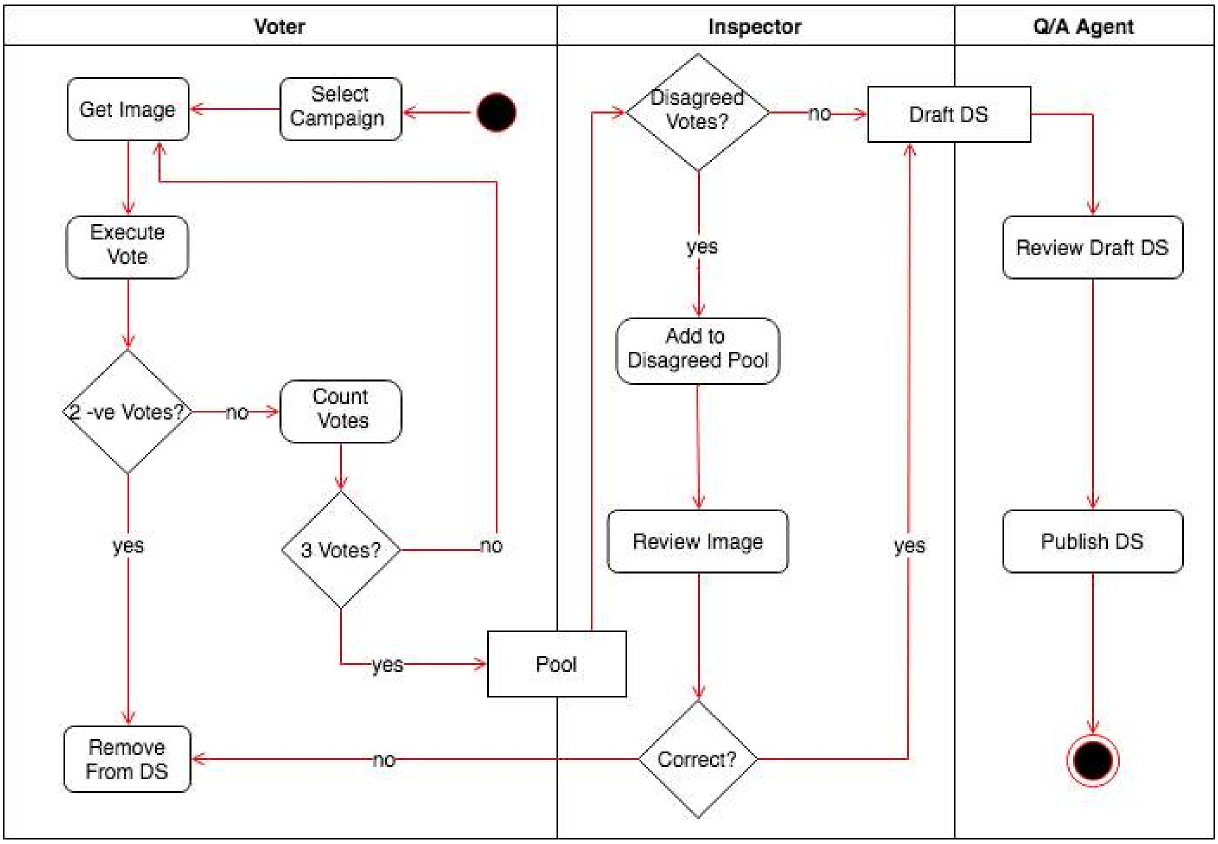


Figure 4.9: UML Activity Diagram of VisAGe GUI

which stores information about the curated dataset. The MD5 field aids the download to make sure it has not been corrupted. The data set is an aggregate that contains all the positive voted files. Several aggregates can also be seen; the campaign collection contains two aggregates: positive and disagreed votes. The image_created entity contains azure_service aggregate, campaign_transition has a set of users and finally, the entity voting comprises of several collections that have the campaign aggregate in each one of them.

## 4.4 Dlib Contour Artistic Approach for Facial Image Pre-processing

### 4.4.1 Introduction

The dlib Contour Artistic (DCA) approach is a facial pre-processing technique that implements facial detection with extra hairline landmark points and crops frontal faces eliminating background noise and unnecessary features. The face detector is similar to the conventional facial detectors discussed in Section 3.1.3. Nevertheless, the hairline

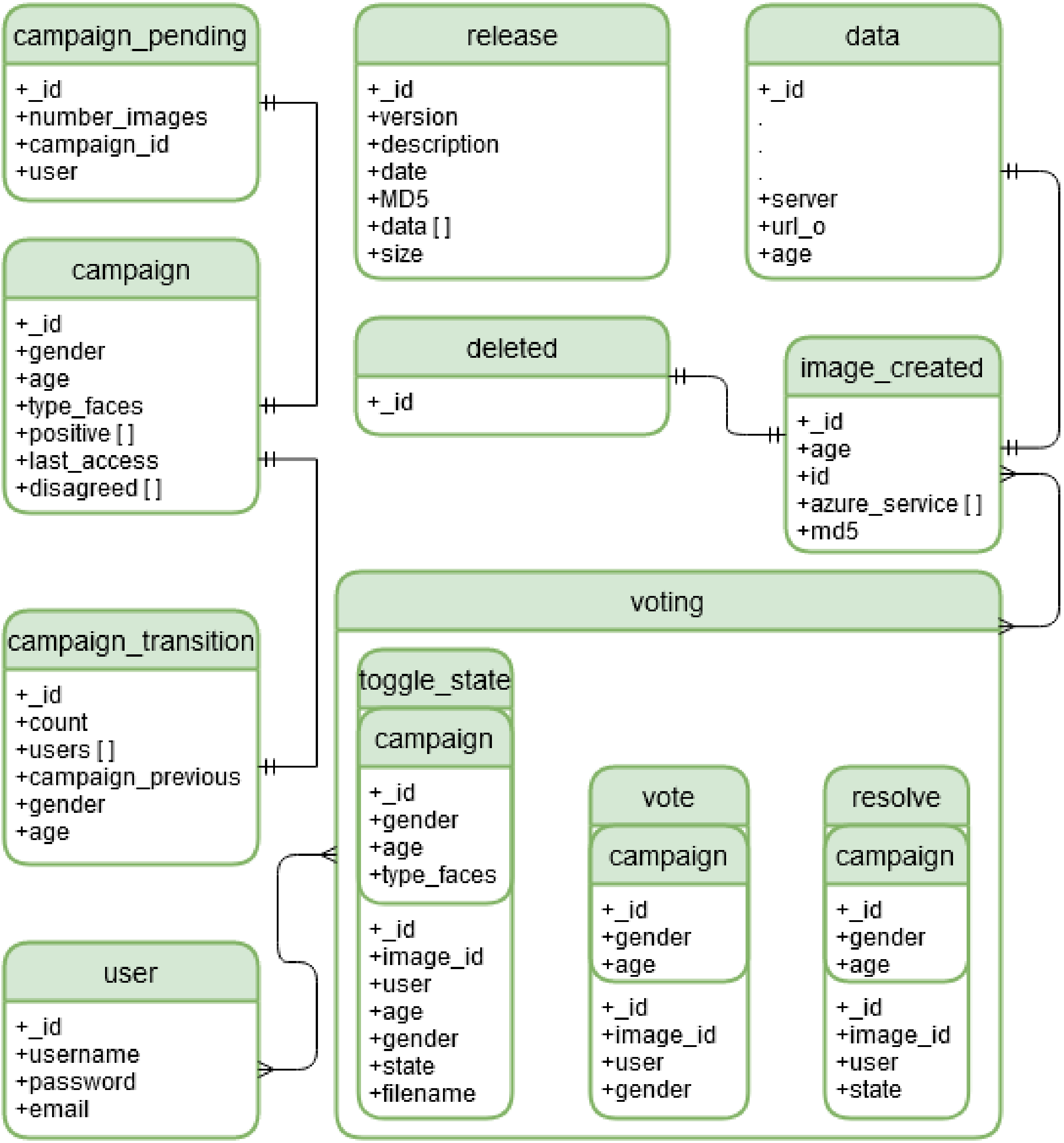


Figure 4.10: VisAGe Database Diagram

is predicted with a facial proportion artistic approach inspired in the works of Andrew Loomis. A build research methodology was applied and the final product was represented by an artefact that is the algorithm for facial age pre-processing.

### 4.4.2 Facial Image Pre-processing

Pre-processing refers to the application of transformations on an image to lessen features and increase the training performance. Both alignment and cropping are pre-processing techniques widely used for face recognition. An important step for an image classification task is to filter unnecessary features that would affect the learning process of a ML algorithm. Facial landmark detection can assist the filtering by being able to select the face contour that will be cropped. Initially, the `Face++` facial landmark detection approach was selected and 1000 landmarks were recognised. The face landmarks can be seen illustrated in Figure 4.11. There were concerns with sending facial images to private cloud servers outside the European Union so an offline approach was favoured. Nevertheless, `dlib` does not support the detection of the hairline which is an important aspect to detect faces due to the presence of wrinkles and other features that a model learns when being trained for age estimation.

Once the landmarks were collected, the left and right eye centre values were processed to further compute the angle between the eye centroids. Next, the median point between the two eyes in the input image was computed and subsequently rotated. Finally an *affine transformation* was applied to the image with warping using the specified matrix, as per Equation 4.2:

$$dst(X,Y) = src(M_{11}x + M_{12}y + M_{13}, M_{21}x + M_{22}y + M_{23}) \tag{4.2}$$

Another approach would be to solve the *procrustes problem* [94] by subtracting centroids, scaling by the standard deviation, and then using the singular value decomposition to calculate the rotation. Once the face was aligned, the new facial landmark positions were detected and a mask for the 273 contour points provided by Face++ was created. The mask is overlaid and the face is cropped. An example of a `dlib` cropped face vs a Face++ cropped face can be seen in Figure 4.11. The `dlib` landmark tool extracts 27 relevant contour points from the 68 points in total against 273/1,000 from Face++.

The major drawback of using the Face++ API is that the images must be sent to a remote cloud service and the accessibility is limited to this cloud environment. In dealing with CSEM investigations, LEAs cannot transmit this sensitive information to a third party service. To overcome this issue, a customised facial cropping technique was implemented using `dlib` as a base and extending it to predict the hairline with a facial proportion artistic approach by Loomis [163].

### 4.4.3 DCA Facial Proportion Artistic Approach

The Face++ and `dlib` pre-processing techniques were analysed to decide which approach was most ideal to be used for the validation of the model. Exploration of the

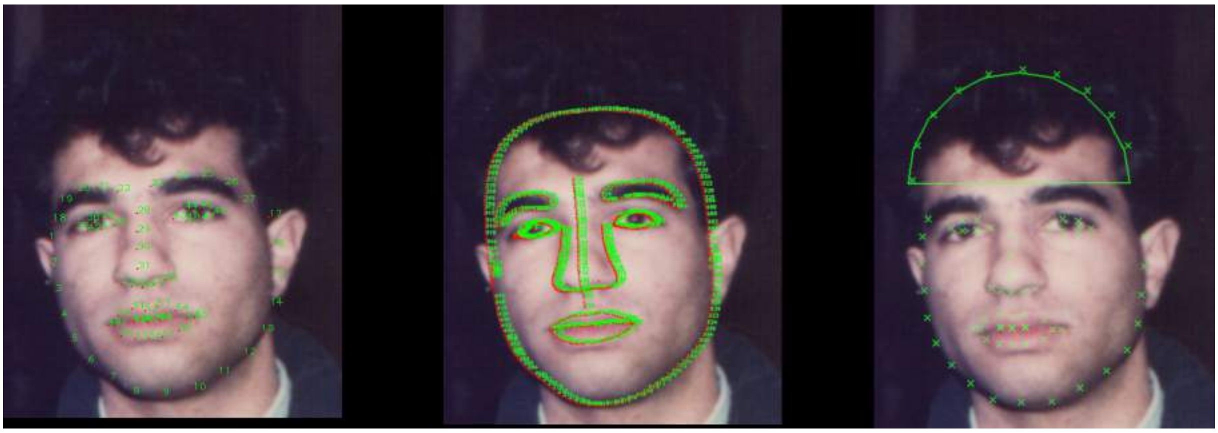


Figure 4.11: Facial Detection Approaches - Image Taken from the FG-NET Aging Database [253] with 64 `dlib` Landmarks (left), 1000 Face++ landmarks (middle), and the DCA approach (right)

Face++ 1,000 landmark points detection tool has 273 contour points (refer to the image on the right of Figure 4.11). Due to the problem with remote cloud services stated in Section 4.4.2 likely being insurmountable for CSEM investigation, the `dlib` tool was selected. The novel pre-processing technique, DCA, was implemented instead based on `dlib` and the Andrew Loomis face proportion technique. `Dlib` returns 68 landmark points from which 27 correspond to the face contour. On the left of Figure 4.11, the `dlib` jawline contour highlighted in green from point number 1 to 27 can be seen. Portions of the head, such as the forehead or wrinkles, are important features that impact the age estimation of a subject; these features are not supported with the 68 landmark detector. The DCA approach addresses this limitation, as can be seen on the right of Figure 4.11). It uses facial proportionality to reconstruct the face and obtain landmarks that are close to the hairline.

Figure 4.12 depicts the proportionality between the nose, eyes and hairline contour. To predict the hairline landmarks, the following steps that emulate the face drawing methodology was carried out as follows:

1. The `dlib` landmark detector is employed to obtain the coordinates $x$, $y$ of the lowest point of the nose which corresponds to the point $N = (x_{34}, y_{34})$.
2. The average distance between the point $N = (x_{34}, y_{34})$ and the intersecting points ($[p_{left}, p_{right}]$) that lie close to a perpendicular drawn from the nose point $N$ towards the contours are computed. The square side is equal to twice this value.

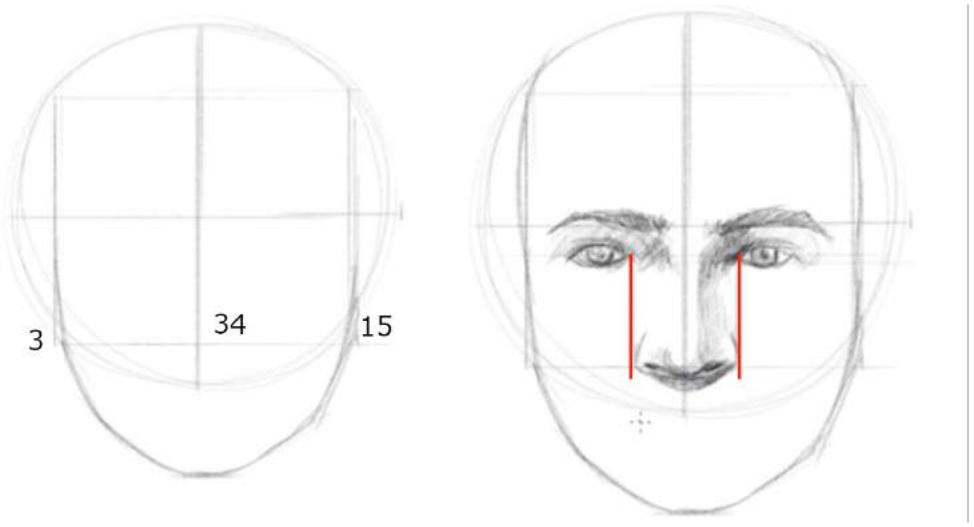


Figure 4.12: Dlib landmark points (3, 34, 15) superimposed over a Loomis face proportion approach sketch [84] (image reproduced with permission)

$$d_1 = \sqrt{(p_{left_x} - p_{34_x})^2 + (p_{left_y} - p_{34_y})^2}$$
$$d_2 = \sqrt{(p_{right_x} - p_{34_x})^2 + (p_{right_y} - p_{34_y})^2}$$

$$square_{side} = \frac{d_1 + d_2}{\cancel{2}} * \cancel{2}$$
$$square_{side} = d_1 + d_2$$

3. Both vertexes $v_1, v_2$ of the square are located and employed to draw the shape within the circle as shown in Figure 4.12.

4. From the centre points $c_x, c_y$, a regular polygon of size *N = 20* (Icosagon) is drawn. Notice that the circle drawn in Figure 4.12 corresponds to the half icosagon drawn in Figure 5.5.

Once the new points are generated for the hairline, it is possible to crop the face by merging both the landmarks obtained from dlib and the aforementioned points. In Figure 4.13 the DCA process can be seen depicted. The input is a raw image and the output produces an aligned cropped image.

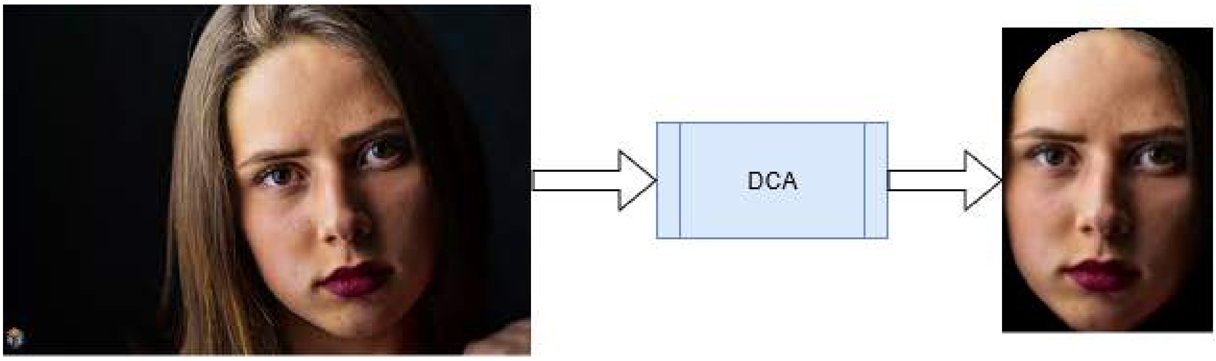


Figure 4.13: Dlib Contour Artistic Approach Process

## 4.5 Underage Age Estimation Model Design

Once a dataset has been curated and ready in production, a model to solve a specific problem can be designed. One of the most critical aspects for defining a model is the number of inputs and the number of outputs. For underage age estimation, the number of inputs would usually be defined by the pixels of the image i.e., 224 x 224 pixels. And the output could be either a multi-class value or a binary class representing children and adults. The models should be trained on a balanced dataset, it is acceptable to mix gender in the dataset. But in other cases, the gender would be segregated into male and female classes. This would demand more effort in the data collection and would result in a more complex problem. It would be convenient to divide the model into an easier problem for an age and gender specific group.

The models presented are the ones developed to attempt to improve the performance of underage age estimation: DS13K, DeepUAge and Vec2UAge. It should be noticed that each model was trained on a different dataset which is discussed in each dataset curation section per model.

### 4.5.1 DS13K

The DS13K model was one of the initial models created in early stages of the research. It was developed based on the architecture of the DEX model. The weakness of the algorithms specifically in the borderline has also been a motivation for the creation of this approach. An ensemble technique that improves the accuracy of underage estimation in conjunction with the DL model (DS13K) has been developed. DS13K has been fine-tuned on the DEX model.

#### 4.5.1.1 Dataset Curation

In order to perform unbiased experimentation with the four offline and cloud-based age estimation services (AWS Rekognition, Azure API, How-Old.net and DEX), it was necessary to construct a balanced dataset. Thus, an equal number of image collection was assured, for each age. The dataset generator proposed in Section 4.6 was used. From Figure 4.3, it can be seen that the relative scarcity of images of young children meant that the number of images per age for the balanced dataset was greatly reduced. The lack of images particularly for ages below 7 can be seen depicted in the aforementioned figure.

The focus of this model was to address the complexity of predicting ages in the boundary between minority and adulthood. Therefore, older ages were not considered. Thus, the dataset was limited to an age range of 0 to 25 inclusive (26 classes). The reason the upper boundary age is 25 is presented in Section 4.5.1.5. For this dataset, 492 images per age (26 classes) were collected. For younger ages, this quantity of images was not available in existing public dataset, requiring the incorporation of additional manually discovered images. This was achieved by collecting images from Flickr[11]. Only photos that were available under an appropriate CC or Public Domain license, and for which accurate age and gender information were available, were considered. The latter information was taken from metadata, such as photo titles, descriptions, or tags. Other images were included from the UTKFace Dataset. IMDB and WIKI photos were avoided but still used in a low proportion. Initially a collection of 15,000 images were gathered but due to non-face recognition, the figure decreased and in order to maintain a balanced dataset, the images had to be reduced to the 492 images per class previously mentioned. Hence, the dataset was limited to a total size of 12,792. To supervise the input of the model, each age class was split into two and the average faces were calculated. A sample of average faces between 16 to 17 year-old subjects can be seen depicted in Figure 4.14. This average strategy was useful to detect occlusion and poor exposure in each bin.

#### 4.5.1.2 Dataset Pre-Processing

Each image is a single frontal face that was cropped and aligned with `dlib` with a dimension of 224 x 224 pixels. Each image was processed by a face detector either by the `dlib` libraries by using histogram-oriented gradient (HOG) or CNN, or the face detection provided by each service discussed in Section 4.1.

[11] Appropriate ethical approval was awarded for this data gathering process. Refer to Appendix A.1

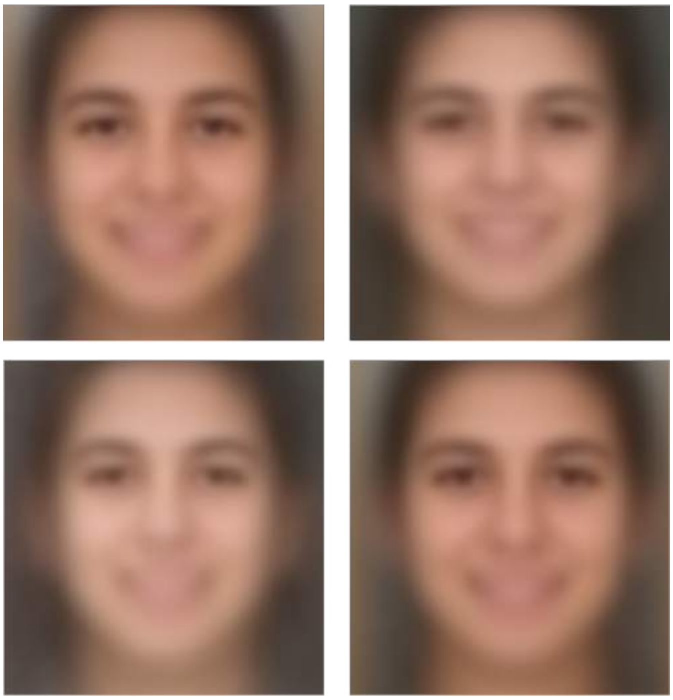


Figure 4.14: DS13K Average Faces between 16 to 17 Year Old's.

The mean image of the dataset was computed. This is a data normalization technique which means that the new mean of all the images will be zero, a pre-processing step to minimize the cost function and therefore optimize the process.

#### 4.5.1.3 Dataset Split

The 12,792 images were divided into 80% for training and 20% for testing.

#### 4.5.1.4 Dataset Augmentation

Flipping images horizontally was the only dataset augmentation technique used to double the amount of images.

#### 4.5.1.5 Network Architecture

The *DS13K* model was fine-tuned on DEX to take advantage of the preexisting layer weights. The previously-mentioned DEX model was built on a VGG-16 architecture as the one depicted in Figure 2.7 with $K = 5$. For the development of the model, transfer learning was used.

The estimation was addressed as a classification problem (Multi-class classifier). The images were grouped into $K = 5$ different classes: 0-5, 6-10, 11-15, 16-17 and 18-25. The ranges were adapted from the "Criminal networks involved in the trafficking and exploitation of underage victims in the European Union" 2018 report [72]. Europol stated that the classification of subjects into one of these age ranges is sufficient. Moreover, the precise age estimation is not crucial for investigators.

#### 4.5.1.6 Hyper-Parameters

Smaller number of iterations produce better results. This might mean that by iterating over the same data many times the model ends up fitting not just the data but also the noise, i.e. over-fitting. 2,000 and 10,000 iterations were selected.

### 4.5.2 DeepUAge

DeepUAge is an underage facial age estimation DCNN model based on a residual neural network of 50 layers (ResNet50). The model is trained on images that were pre-processed with the DCA approach technique discussed in Section 4.4.3.

#### 4.5.2.1 Dataset Curation

Age and gender estimation models require a large number of images with real age and gender labels. Moreover, the data must be balanced within each class and this represents a challenge. Nevertheless, it was possible to build a balanced set divided with images from two sources: the VisAGe dataset supplemented by the dataset generator discussed in Section 6.1.1.

The creation of VisAGe involved the annotating of the largest set of underage images to date. These images are CC licensed with initially indicative age gathered from Flickr. Each of these photos were processed by face and gender detection algorithms, and other associated metadata were compiled such as dimensions, title, tags, comments, dates, etc. Given the level of error rates in automated facial age estimation and gender identification, each of these images were subjected to human age and gender verification. Each photo was voted on by three human assessors and if the decisions on age and gender were unanimous, the photo was added to the dataset. The set used in this model consists of 19,446 images from the age range 1 to 18. Further detail regarding age and gender per class is depicted in Table 4.4.

It is notable that the age ranges contain an unbalanced amount of images within each age and/or gender group. The average number of male images per age is 521 versus 557 females.

| Age | Combined | Male | Female |
|---|---|---|---|
| 1 | 4,236 | 2,292 | 1,944 |
| 2 | 2,722 | 1,485 | 1,237 |
| 3 | 2,280 | 1,071 | 1,209 |
| 4 | 2,434 | 1,110 | 1,324 |
| 5 | 1,227 | 515 | 712 |
| 6 | 984 | 462 | 522 |
| 7 | 974 | 418 | 556 |
| 8 | 686 | 315 | 371 |
| 9 | 453 | 256 | 197 |
| 10 | 401 | 217 | 184 |
| 11 | 371 | 154 | 217 |
| 12 | 211 | 103 | 108 |
| 13 | 354 | 171 | 183 |
| 14 | 217 | 142 | 75 |
| 15 | 337 | 91 | 246 |
| 16 | 589 | 184 | 405 |
| 17 | 285 | 204 | 81 |
| 18 | 660 | 193 | 467 |
| **Total** | 19,446 | | |

Table 4.4: VisAGe Dataset Demographics - Facial Images per Class per Gender from 1 to 18 Year-Old Subjects

To monitor the inputs and present a sample of the dataset, an average image per class was calculated. Average faces from age 1 to 18 can be seen in Figure 4.15.

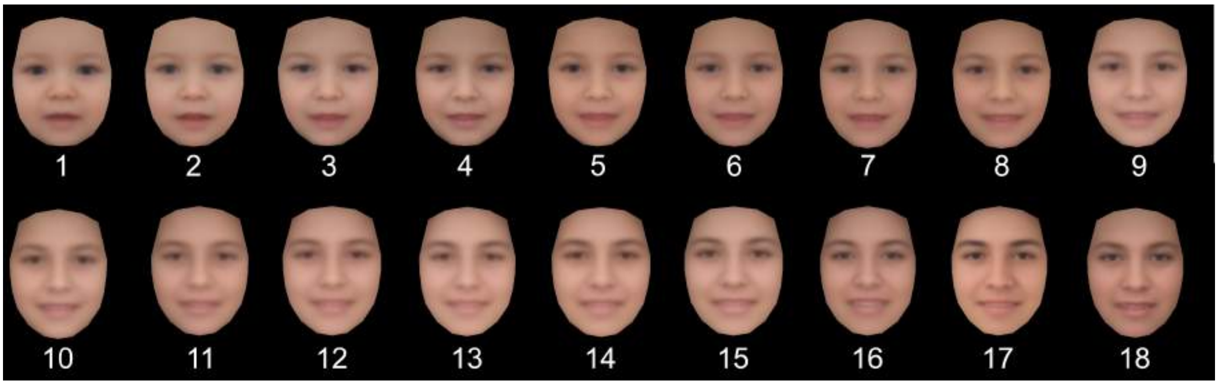


Figure 4.15: VisAGe Dataset - Average Face per Age from 1 Year-Old (Top-left) to 18 Year-Old (Bottom-right) Subjects.

The face cropping technique used in this model is also applied and visible on each face.

The age estimation model was trained with the majority of images from the VisAGe dataset and was prepared in a balanced fashion. Due to the size of the dataset available, 800 photos were selected from each class. When enough images were available for each class, a balanced amount of images were selected for both male and females. In the age classed of 8 and higher, there were insufficient images to fulfil the 800 training/testing images. As a result, the remainder of the 800 images used were filled with the Anda et al. [6] facial age dataset generator with randomly obtained images from different datasets including FGNET, IMDB-WIKI, FERET, MEDS.

Although the model application is for underage images, it was necessary to consider 2 additional years over the maximum year limit (18 years old). Since the best performance of existing approaches are approximately 2-3 MAE in years, the chosen limit was 20. Furthermore, the additional age classes were also completed with the aforementioned dataset generator.

#### 4.5.2.2 Dataset Pre-Processing

The input image size for width and height chosen was 224 with the depth set to 3. These values were chosen due to the hardware limitations and the decreased performance experienced with images of smaller dimensions. The DCA approach discussed in Section 4.4.3.

#### 4.5.2.3 Dataset Split

A balanced dataset of 16,000 images was prepared. An 80% of the images were used for training and 20% were used for validation. The model was further tested with 1,000 additional images that were gathered from the UTKFace dataset explained in Section 3.3.3.2 and the aforementioned dataset generator. These images amounted to 50 images per class. The labelled ages from the former dataset were predicted with the DEX model and validated by the researchers [270]. The ages of the latter dataset were obtained from each source that provided them (Refer to Section 4.2).

#### 4.5.2.4 Dataset Augmentation

Data augmentation techniques such as horizontal flip, left and right rotations, random zoom improvement, stochastic distortion, random colour, contrast and brightness with a specific minimum and maximum factor, and random erasing with a fixed area were used.

#### 4.5.2.5 Network Architecture

The proposed method is pre-trained on the ImageNet Dataset. The last FC softmax layer with 1,000 outputs has been replaced by an FC-softmax activation function layer of 20 outputs that correspond to the age classes studied (1 to 20 years old) to suit our needs. Subsequently, the parameters of the convolutional layers during the training process have been frozen. The ResNet50 architecture employed for facial age estimation and the replacement in the last layer with 20 outputs can be seen in Figure 4.16. The age estimation problem was treated as a classification task and therefore, a categorical cross-entropy logarithmic loss function was used.

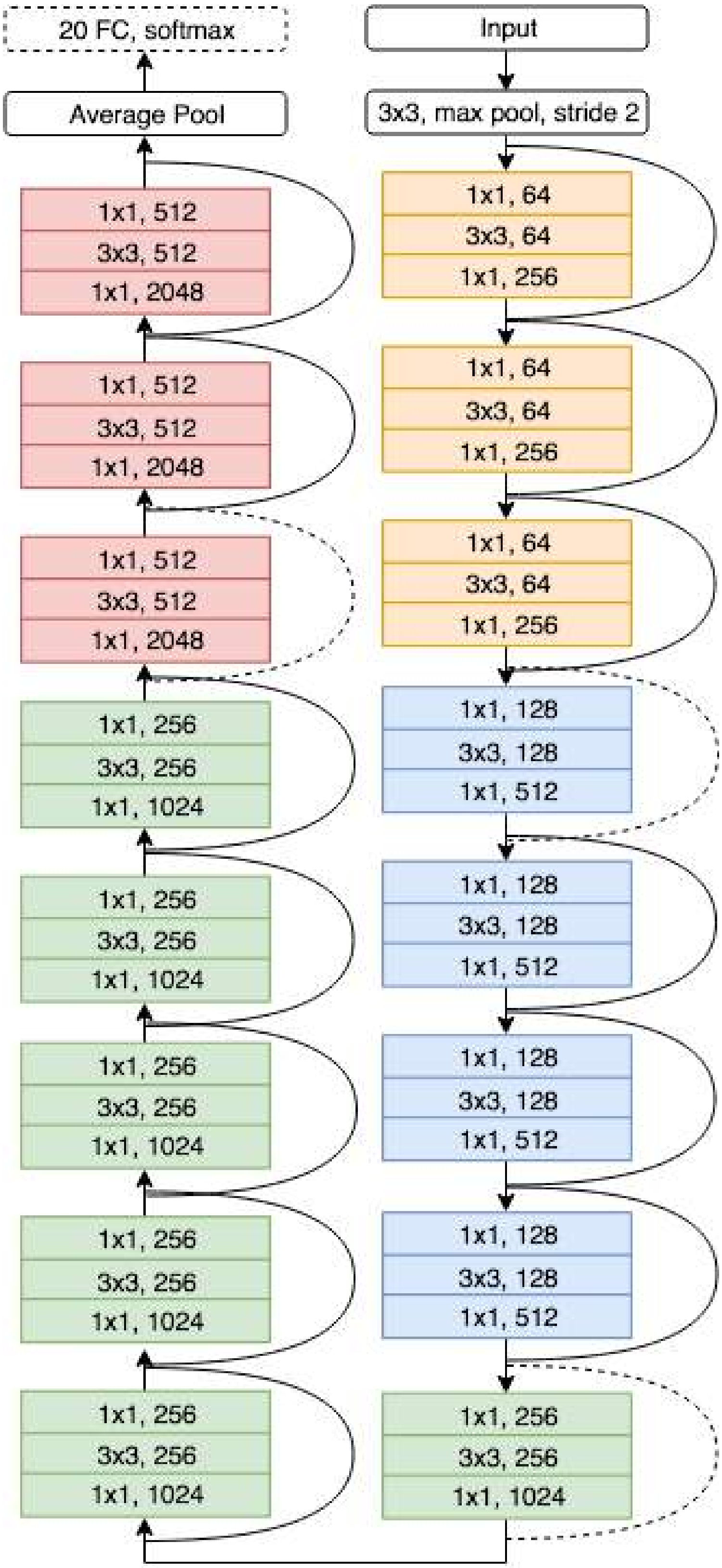


Figure 4.16: ResNet50 Pre-trained on ImageNet with 20 Outputs in the FC Softmax Layer. Grouping of Convolution Layers are Denoted by Colour.

#### 4.5.2.6 Hyper-parameters

The batch size is a hyper-parameter that defines the number of training samples used in one iteration and is directly proportional to the memory space required. A batch size of 64 was chosen due to random access memory (RAM) limitations with the development server. One forward pass and one backward pass of all the training examples are referred to as epochs. A reference of 100 epochs was chosen. However, the training process was monitored and an early stopping implementation to prevent over-fitting was accomplished. The metric used for accuracy was the MAE, i.e., the average of the absolute mean error between the ground truth and the predicted value. In the experiments, the MSE and MAE are used as performance metrics but only the MAE is reported.

The optimiser chosen was SGD that includes support for momentum, learning rate decay, Nesterov momentum. SGD demonstrates excellent performance for large-scale problems [30]. The LR is a hyper-parameter that controls the number of changes affected by the model in response to the estimated error each time the model weights are updated. The selected value for the initialiser was 0.1 and the momentum was 0.9. The latter is a parameter that accelerates SGD in the relevant direction and reduces oscillations.

### 4.5.3 Vec2UAge

Vec2UAge is a novel regression-based model that uses facial embeddings from FaceNet as feature vectors for training, from the VisAGe and Selfie-FV datasets. The dataset and age determination is specified in Section 4.3.2 and Section 3.3.4.1 respectively. A balanced, unbiased dataset was created for testing and validation. Data augmentation techniques were evaluated to further be used to expand the training dataset. The LR is one of the most important hyper-parameters for DNN. A cyclic LR approach was used to find the optimal initial value for *LR* and the performance was evaluated.

#### 4.5.3.1 Dataset Curation

The size of the dataset required is dictated by both the complexity of the problem that is trying to be solved and the quality of the images. The proposed facial age dataset is a merge between the underage group range pertaining to VisAGe and Selfie-FV. The counts per age distribution of VisAGe, Selfie-FV and a combination of both can be seen in Figure 4.17. The number of underage subjects used for this section of research is 7,419 that belong to the age group of 8 to 18 year old's. The contribution of these images for the combined dataset can be seen depicted in the aforementioned figure.

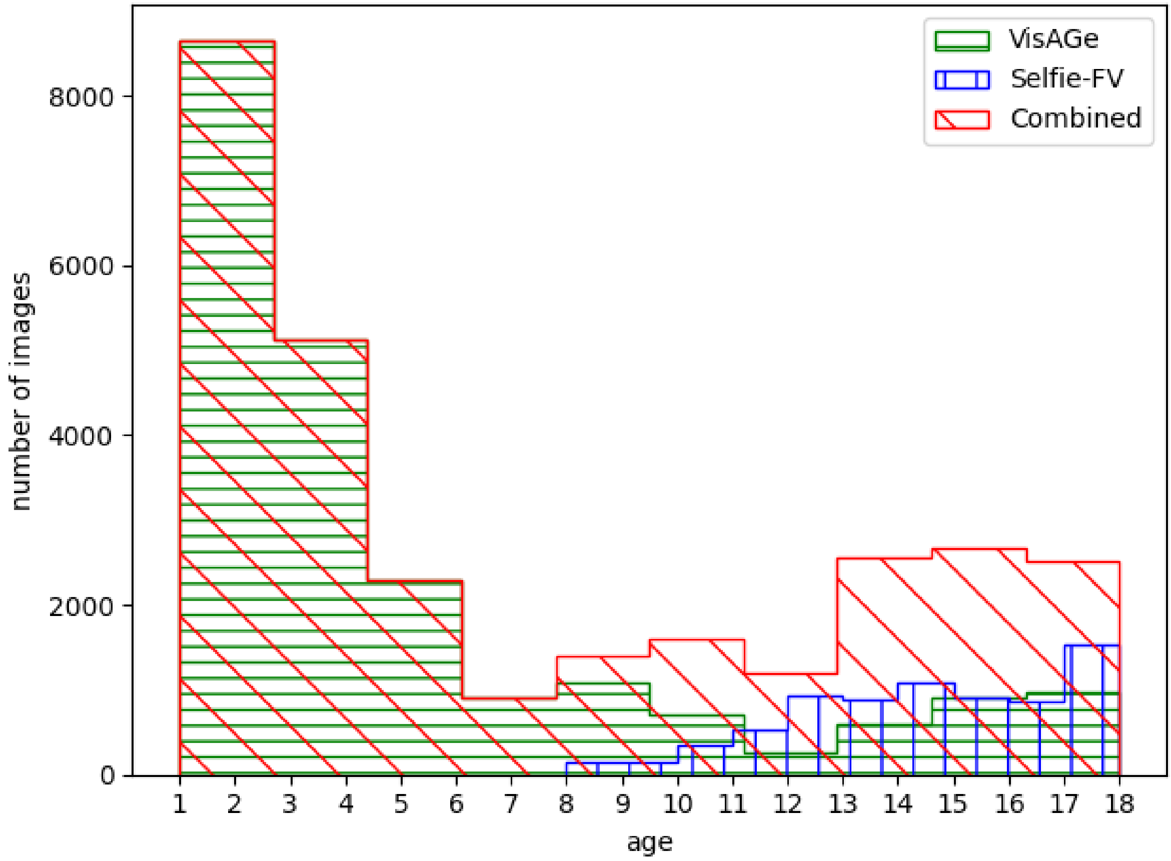


Figure 4.17: Histogram of Age Count Distribution: VisAGe, Selfie-FV and Combined

#### 4.5.3.2 Dataset Pre-Processing

The images for the selected datasets have already been curated and are predominately frontal face photographs of a single subject. Exposure, occlusion, noise and emotion are influencing factors on the accuracy of underage facial age estimation [9]. As a result, images have been discarded according to the level of these factors (with exception of emotion that has a minuscule to strong impact on the accuracy depending on the age and type of emotion as per Section 6.4) thus decreasing the problem to a smaller one.

Face detection is usually needed for age estimation; while reducing the number of pixels to be evaluated, unwanted background and noise is also addressed. Facial detection was employed prior to the facial embedding extraction. This step assured that a face was present so that the succeeding step would not fail. Once the face has been recognised, it is cropped to the detected face rectangle and resized to a size of 224 x 224 pixels.

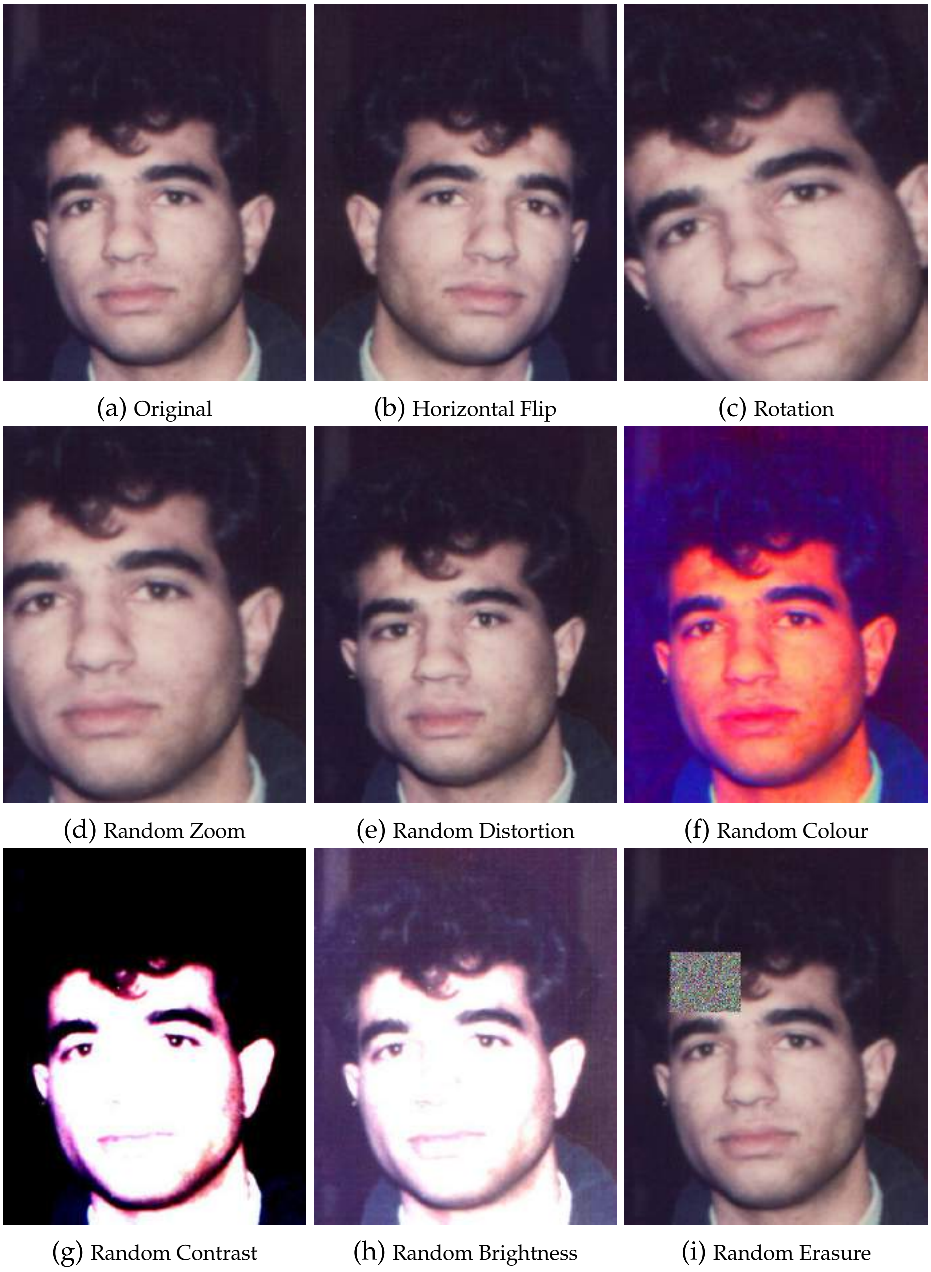
#### 4.5.3.3 Dataset Split

It is noticeable in Figure 4.17 that there is a decrease in the amount of images in the 6 to 8 and 11 to 13 bins. Nevertheless, an unbiased balanced testing/validation dataset was obtained with 500 images per class, leading to a non-augmented dataset of 9,000 images that can be used both for validation and testing or simply for validation.

Stratified Shuffle Split [200] was applied to divide the dataset in validation and test where the test dataset was 50% of the validation set. The shuffling technique applies stratified randomised folds – made by preserving the percentage of samples for each class.

#### 4.5.3.4 Data Augmentation

The training set was the only one affected by data augmentation. The data augmentation techniques used are both geometric and photometric transformations; horizontal flip, rotation, random zoom, random distortion, random colour, random contrast, random brightness and random erasing were applied. The visual effects of the several data augmentation techniques can be seen depicted in Figure 4.18.


(a) Original (b) Horizontal Flip (c) Rotation

(d) Random Zoom (e) Random Distortion (f) Random Colour

(g) Random Contrast (h) Random Brightness (i) Random Erasure

Figure 4.18: Facial Image Augmentation Techniques: Original Image Taken from FG-NET Aging Database [253]

#### 4.5.3.5 Network Architecture

Having the face vectors calculated previously, the inputs rather than being pixels, are float values that can be processed with a much simpler neural network. A simple 4-layer neural network with 512, 256, 128, 1 units at each layer was constructed as shown in Figure 4.19. This network configuration was the result of the use of an optimisation framework.

For every hidden layer, a ReLU activation function was used. The input layer is not

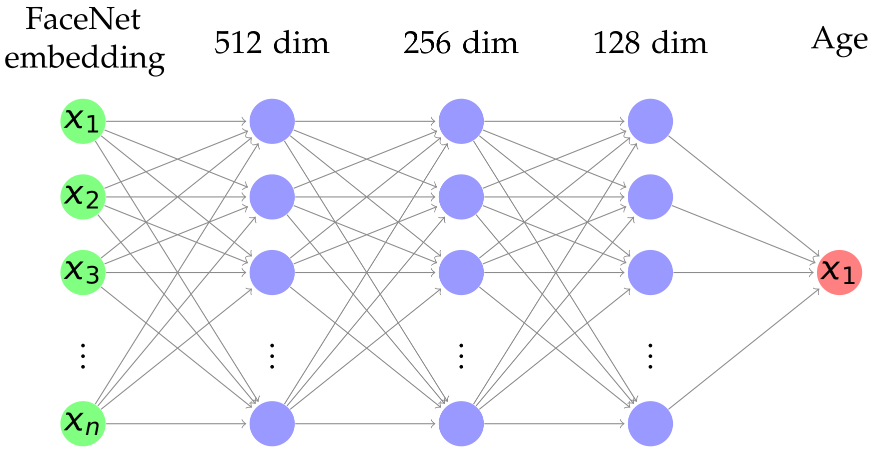


Figure 4.19: Proposed neural network: for every hidden layer a ReLU activation is used

considered a layer of neurons, but rather the entry of the facial embeddings of size 512. The first hidden layer consists of 512 neurons, followed by the next layer which consists of 256 neurons and the 3rd layer is of size 128 and will generate the final results. The last layer consists of a single output neuron as an age regressor. ReLU is one of the most commonly used activation functions in neural networks. It relies on a simple calculation that returns the input if the value is greater than 0 – otherwise it returns 0. The function can be seen in Equation 4.3.

$$f(x) = max(0, x) \tag{4.3}$$

#### 4.5.3.6 Hyper-Parameters

The optimisation algorithms chosen were ADAM, ADAGRAD, SGD and SWA. Both fixed and dynamic LR approaches were considered. A cyclic LR approach was used to find the optimal initial value for LR. Lastly, the number of epochs selected was 100, but early stopping was implemented meaning that it will stop training once the model performance stops improving on the hold out validation dataset.

## 4.6 Influencing Factors on the Accuracy of Automated Underage Facial Age Estimation

An introduction to the influencing factors was previously discussed in Section 3.2.11. In this section, we describe the methodology on how the assessment of the influencing factors was accomplished with the VisAGe dataset through the employment of cloud-based services such as Microsoft Azure and AWS Amazon Rekogntion. Both services were introduced previously and have facial detection capabilities, age estimation and facial attribute detection.

The VisAGe dataset was processed by Azure's Face API and AWS Rekognition and the age estimations obtained from the two cloud services were measured against the ground-truth age in the dataset. The absolute difference between the two values has been denoted as the error difference ($\mathrm{Er_d}$) and can be seen depicted in Equation 4.4 where $y$ is the *machine predicted age* and $x$ is the *ground-truth age*.

$$Er_{\mathrm{d}} = |y - x| \tag{4.4}$$

Equation 4.4 has been used as the principle measurement in assessing the accuracy of the underage facial age estimation. Additional features of both cloud facial analysis services were utilised to classify and annotate the data as per Tables 4.5 and 4.6. To process the correlations between variables, the object attributes have been broken down into categorical values.

Having determined the attributes of each image and their associated $\mathrm{Er_d}$, the correlation between the two variables of data was then calculated to identify which attributes were the larger influencing factors of $\mathrm{Er_d}$ and by what gravity, e.g., weak, mild, or strong.

Attributes with mild to strong correlations had influence in the accuracy of the underage facial age estimation. Henceforth, the investigation has been split into the gravity of errors in order to identify traits that most of the data adhere to, versus the traits of the minorities, i.e., data that lies within $\mathrm{Er_d} > 5$.

### 4.6.1 Cloud Services

Two cloud services were used in this section of study to provide the underage facial age estimations of each image within the VisAGe single-faced dataset; Amazon AWS Rekognition Service and the Microsoft Azure Face API service.

#### 4.6.1.1 Microsoft Azure: Face API

This service assisted the annotation of each record according to the detected facial attributes such as perceived emotion, presence of facial hair and makeup, facial expressions like happiness, contempt, neutrality, and fear, etc. A comprehensive list is presented in Table 4.5.

Table 4.5: Microsoft Azure Cognitive Services Face API Attributes [175].

| Field | Description |
|---|---|
| emotion | Neutral, anger, contempt, disgust, fear, happiness, sadness, and surprise. |
| noise | Noise level of face pixels. |
| age | “Visual age” number in years. |
| gender | Estimated gender with male or female values. |
| makeup | Presence of lip and eye makeup. |
| accessories | Accessories around face, including ‘headwear’, ‘glasses’ and ‘mask’. |
| facialHair | Moustache, beard and sideburns. |
| hair | Group of hair values indicating whether the hair is visible, bald, and hair colour if hair is visible. |
| headPose | 3-D roll/yaw/pitch angles for face direction. |
| blur | Face is blurry or not. ‘Low’, ‘Medium’ or ‘High’. |
| smile | Smile intensity, a number between [0,1]. |
| exposure | Face exposure level. Level returns ‘GoodExposure’, ‘OverExposure’ or ‘UnderExposure’. |
| occlusion | Values are Booleans and include ‘foreheadOccluded’, ‘mouthOccluded’ and ‘eyeOccluded’. |
| glasses | Glasses type. Values include ‘NoGlasses’, ‘ReadingGlasses’, ‘Sunglasses’, ‘SwimmingGoggles’. |

#### 4.6.1.2 Amazon AWS: Rekognition Service

Amazon Rekognition is a pre-trained image analysis service. Its face detection and analysis service was used to perform several visual analyses on VisAGe; extracting facial attributes such as facial hair, expressions, etc., detected on each single-faced image. The attributes, as outlined in Table 4.6, were then correlated against Amazon’s facial age estimator to provide a comprehensive evaluation on the accuracy of underage facial age estimation against the influencing factors.

### 4.6.2 Pearson Correlation Coefficient

The Pearson correlation coefficient (PCC) measures the linear correlation between two variables. In this work, these are the attribute and $\mathrm{Er_d}$. The value of the coefficient lies between +1 and -1; where ±1 indicates a perfect correlation and 0 represents no

Table 4.6: Amazon AWS Rekognition Attributes [17]

| Field | Description |
|---|---|
| Age.Range | Estimated age range. |
| Smile.Value | Smile value detected true or false. |
| Eyeglasses.Value | Eyeglasses detected true or false. |
| Sunglasses.Value | Sunglasses detected true or false. |
| Gender.Value | detected gender on subject. |
| Beard.Value | Beard detected true or false. |
| Moustache.Value | Moustache detected true or false. |
| EyesOpen.Value | Open eyes detected true or false. |
| MouthOpen.Value | Open mouth detected true or false. |
| Emotions | Detection true or false for each array. |
| Landmarks[0] | X-axis and Y-axis positions. |
| Roll (Degree) | Face titled to the side. |
| Yaw (Degree) | Face turned to the side. |
| Pitch (Degree) | Face titled up or down. |
| Brightness | Brightness of the image. |
| Sharpness | Sharpness of the image. |
| Confidence | Certainty of the estimation. |

correlation at all. A negative coefficient signifies an inverse relationship between the variables. For a sample of data, such as that examined here, the PCC is often represented as $r_{xy}$ and is defined in Equation 4.5:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.5}$$

where $n$ is the size of the sample, $x_i$ and $y_i$ are individual sample pairs and $\bar{x}$ and $\bar{y}$ are the mean of $x$ and $y$. The correlation value obtained for each sample, i.e., the facial attribute and $\mathrm{Er_d}$ pair, was matched inline with a scale of weak, mild, or high. It is important to note that for the purpose of this work, weak, mild and strong correlations are characterised with 0.1 – 0.29, 0.30 – 0.49, 0.50 – 1 correlation values respectively (whereby the negatives of these values represent inverse correlations). These definitions have been defined in a computer forensic related study regarding analysis of correlations of Internet usage [224]. Conversely, correlation close to zero, specifically within the $-0.1 \;--\; 0.1$ range has been referenced as minuscule correlation.

CHAPTER

FIVE

# IMPLEMENTATION

## 5.1 Dataset Generator

The dataset generator was coded in python 2.7 with a MVC software design pattern. The model layer mainly contains the real-world components; in this case, the databases in general. The controller acts as a liaison between the model and the view, receiving user input and deciding what to do with it. And the view contains the functions that directly interact with the user such as the Delete, Save, Predict and Stats button. The dataset metadata is uploaded to a MongoDB collection, a python package is created for each dataset; in each package, there are methods that enable the upload of the metadata into MongoDB. Furthermore, a class is created per each dataset. The non-relational database is managed with the *pymongo* library. The GUI is based on the *Tkinter* package and the images are manipulated with the *Pillow* bundle. Bitbucket was used for version control and the repository is publicly available in the following URL: `https://bitbucket.org/4nd4/image_database.git`. The last commit was **1015616** on 20 March 2018.

### 5.1.1 Database Superclass

All the data sources inherit from the Database superclass. An overview of the methods can be seen in listing 1.

The *set_filter_list* method initialises the filter list, *get_filter_list* obtains the filtered list. The filter list is composed of images that a user deemed as repeated images, low quality or wrongly labelled images. Once filtered, the images will not appear on the dataset generator GUI again. Following are the getters and setters for path related methods: *get_image_path*, *set_root_path*, etc. Finally, *get_random* returns a random image from a dataset and makes sure that the following instances to the method return unique ran-

**Listing 1** Database Superclass

```
class Database:
    def __init__(self, name):...
    def get_name(self):...
    def get_db(self):...
    def set_filter_list(self, filter_list):...
    def get_filter_list(self):...
    def get_image_path(self):...
    def set_root_path(self, path):...
    def set_crop_path(self, path):...
    def get_root_path(self):...
    def get_crop_path(self):...
    def get_random(self):...
    def get_delete(self):...
    def delete_random_documents(self):...
```

dom images each time. For more information, the software is publicly available.

### 5.1.2 Dataset Plugins

A plug-in modality was implemented due to scalability needs. Facial age labelled datasets are released not so often but the system was designed so it can support the inclusion of new datasets through the development of independent packets that share certain methods in common such as dataset upload, managing paths, counting images, obtaining random images, etc. An example of a dataset module can be seen depicted in listing 2. It is observed that the MORPH class inherits from the Database Superclass discussed in Section 5.1.1.

**Listing 2** Morph Methods

```
class MORPH(Database):
    def __init__(self):...
    def set_image_path(self, record):...
    def upload_database(self, root_dir):...
    def get_filtered_images(self, filtered_list):...
    def get_record(self, age, gender, filtered_images):...
    def count_images(self, age, gender, filtered_images):...
    def get_random_image(self, age, gender, filtered_images):...
    def manage_paths(self, record):...
```

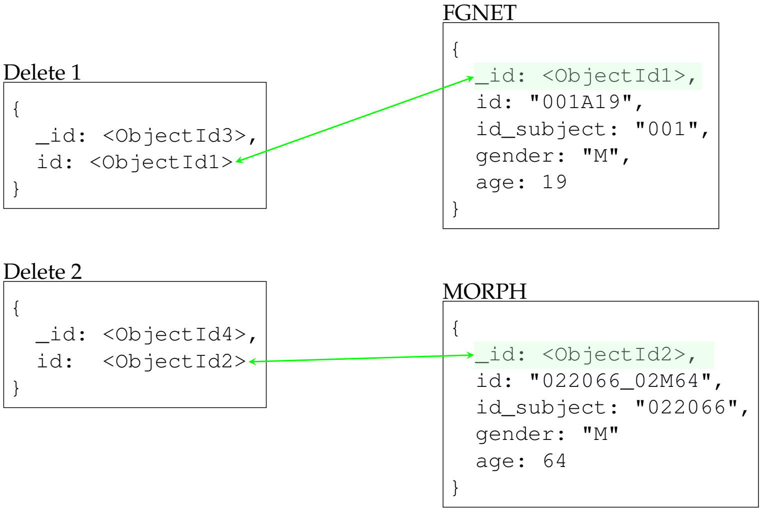
### 5.1.3 NoSQL Dataset Implementation

MongoDB 3.4 was employed and the design used was the one described in Section 4.5. A NoSQL script to insert a record in the FGNET collection can be seen depicted in Listing 3. When the collection does not exist and the script is executed, the collection creates the structure without having to invoke the *db.createCollection* function.

**Listing 3** FGNET Record Insertion Example

```
db.FGNET.insert(
{
    "id": "001A22"
    "id_subject" : "001",
    "gender" : "M",
    "age" : 22
})
```

In the DELETE collection the _id value from all the discarded images of all the datasets are stored. The _id field is the primary key with an ObjectID value so that each document can be uniquely identified in the collection. Once the value is stored in the DELETE collection, the dataset generator will never produce that image again. In Figure 5.1 the JSON documents per collection can be observed of FGNET and MORPH; it is also seen that the _id field of the datasets will be stored in the DELETE collection with a 1-to-1 relationship.



FGNET

```
{
  _id: <ObjectId1>,
  id: "001A19",
  id_subject: "001",
  gender: "M",
  age: 19
}
```

Delete 1

```
{
  _id: <ObjectId3>,
  id: <ObjectId1>
}
```

Delete 2

```
{
  _id: <ObjectId4>,
  id:  <ObjectId2>
}
```

MORPH

```
{
  _id: <ObjectId2>,
  id: "022066_02M64",
  id_subject: "022066",
  gender: "M"
  age: 64
}
```

Figure 5.1: Implementation of MongoDB Collections: FGNET and MORPH Collections and the 1-to-Many relationship with the DELETE collection.

### 5.1.4 Dataset Generator GUI

A key component of the system architecture is the manual filtering step. The amalgamation of various random images per age class generates a dataset with noise that can only be effectively filtered through user interaction. Users are presented with an interface whereby they are able to discard images that are not useful and randomly generate new images by clicking on each image button, as illustrated in Figure 5.2.

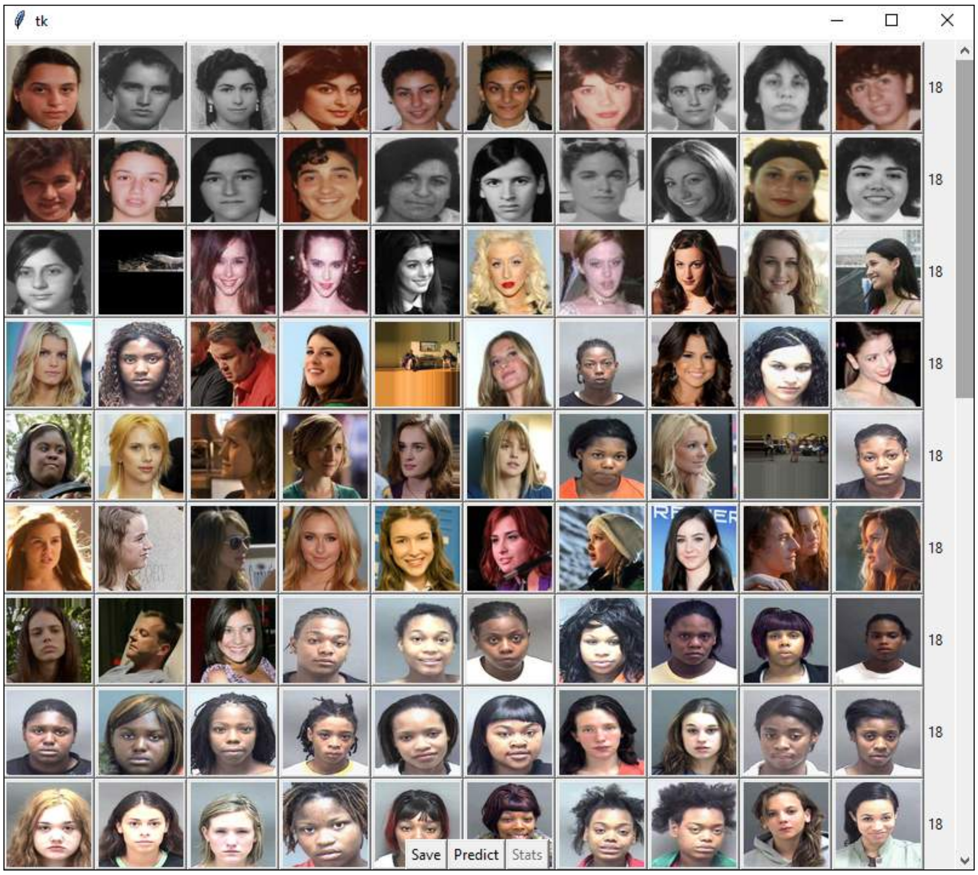


Figure 5.2: Dataset Generator Software

The graphical interface was produced with *Tkinter*. The tkinter package ("Tk interface") is the standard Python interface to the Tk GUI toolkit. Both Tk and tkinter are available on most Unix platforms, as well as on Windows systems.

The "View" folder refers to the view layer and contains several files that manage the GUI; the files are preceded with the "vw" prefix. They manage both the plots and the graphical interface. The file *vw_image.py* contains the entire graphical functionality. The method *drawGUI* of the aforementioned file implements the decorator pattern that was discussed in Section 4.2.2. The image dataset class is passed as a parameter and the decorator extends the functionality of the images enabling the drawing of the photos.

The interface is first executed on a terminal with the command in Listing 4.

**Listing 4** Dataset Generator Execution Command in the Terminal

```
python run.py -g 'male' -min_age 18 -max_age 20 n 100
```

The file is *run.py* and the description of the arguments can be seen depicted in Table 5.1. The software would generate 300 images (100 x 3) from 18 to 20 year-old females and would present the interface seen in Figure 4.6. The user then would have to manually check if the age and gender is consistent with the input parameters, and if the image corresponds to a face. These images are automatically stored in a folder that was configured in the *configuration.py* file, specifically the **path_directory_creation** configuration parameter.

| **argument** | **name** | **description** |
|---|---|---|
| *g* | gender | gender of the subjects |
| *min_age* | minimum age | initial age of the dataset |
| *max_age* | maximum age | age limit of the dataset |
| *n* | number | number of images to produce |

Table 5.1: Arguments of the run.py file of the Dataset Generator

## 5.2 Evaluation Platform

The evaluation platform is the core of the evaluation research, it facilitates the automatic execution of age estimation predictions from the cloud and local based services discussed in Section 4.1.1 and 4.1.2. The services inherit the methods and attributes from the Service Superclass. The service evaluation user interface is a terminal and the command can be seen depicted in Listing 5.

**Listing 5** Evaluation Platform Execution Command in the Terminal

```
python run_evaluation.py
```

The **run_evaluation.py** script requires previous parameter configurations to run; the parameters can be seen depicted in Table 5.2.

### 5.2.1 Service Superclass

All the services inherit from the Service superclass. An overview of the methods can be seen in listing 6.

| parameter | description |
|---|---|
| *path* | path of the subjects |
| *db_list* | list of datasets |
| *service* | list of offline & online services |

Table 5.2: Parameters of the run_evaluation.py file of the Service Evaluation

**Listing 6** Service Superclass

```
from datetime import datetime
from pymongo import MongoClient
from bson.objectid import ObjectId

cl = MongoClient()

class Service:

    def __init__(self, name, limit):...
    def get_path(self):...
    def set_path(self, path):...
    def get_limit(self):...
    def get_response(self):...
    def get_name(self):...
    def get_db(self):...
    def service_evaluated(self, subject_id):...
    def get_records_stats(self, gender=None):...
    def get_records(self):...
    def execute_predictions(self, directory_path):...
    def execute_prediction(self, file_path):...
```

The get and set path methods (*get_path & set_path*) are used to manage the path to the balanced dataset which was created by the dataset generator in Section 5.1. According to Table 4.1, there are limitations per free use for each cloud-based service. As seen in the Service superclass listing previously mentioned, the Service class is initialised with the name of the service and the limit. The *get_limit* is a method that obtains the limit of the service calls. For instance, Microsoft Azure Face API is limited to 30,000 calls per month for free. The evaluation service platform supports the functionality to limit usage per service and once the limit has reached, an email is sent to notify about this warning. A typical cloud API response is returned in a JSON format; the method *get_response* is used to obtain the response from the service.

### 5.2.2 Service Plugins

The plug-in modality implemented for "service" has the same logic as the database plugin which was mentioned in Section 5.1.2. The aim of this logic is to allow the interoperability with future services that would be able to be implemented in a simple fashion. Age estimation models are released often and the system is designed so it can support the inclusion of new age estimation models through the development of independent packets that share certain methods in common such as exception management, subscription key retrieval, response management, real/predicted age retrieval, etc. An example of a service module can be seen depicted in listing 7. It is observed that the Azure class inherits from the Service Superclass discussed in Section 5.2.1.

**Listing 7** Microsoft Azure Methods

```
import http.client, urllib, json
import time

from datetime import datetime

import configuration
from model import db_functions, db_image
from model.db_service import Service
from model.db_image import Database


class Azure(Service):
    def __init__(self):...
    def set_exception(self, exception):...
    def get_exception(self):...
    def get_subscription_key(self):...
    def set_response(self, image_path):...
    def _process_response(self, image_path):...
    def get_real_age_and_predicted(self, images, gender):...
```

### 5.2.3 NoSQL Service Implementation

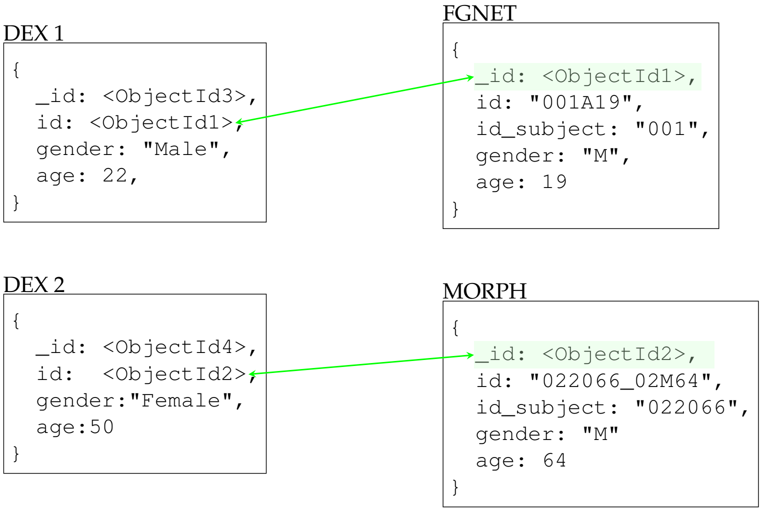
The design used was the one described in Section 4.7. A NoSQL script to insert a record in the SRV_DEX collection can be seen depicted in Listing 8. When the collection does not exist and the script is executed, the collection creates the structure without having to invoke the *db.createCollection* function.

In each DEX Service registry, the field *_id* of the relevant dataset is stored, with the predicted age and gender values. The same applies for every other service. This would

**Listing 8** DEX Record Insertion Example. 12-byte "ObjectId" in line 5 shortened for readability.

```
1  db.SRV_DEX.insert(
2  {
3      "gender" : "Female",
4      "age" : 9,
5      "subject_id" : ObjectId("595256a2...34e6fd48"),
6  })
```

control the evaluation performance procedure to not produce duplicate predictions and henceforth, save time and credits. In Figure 5.3 the JSON documents per collection can be observed of FGNET, MORPH and the DEX documents. it is also seen that the _id field of the datasets will be stored in the DEX documents with a 1-to-1 relationship.



DEX 1

```
{
  _id: <ObjectId3>,
  id: <ObjectId1>,
  gender: "Male",
  age: 22,
}
```

FGNET

```
{
  _id: <ObjectId1>,
  id: "001A19",
  id_subject: "001",
  gender: "M",
  age: 19
}
```

DEX 2

```
{
  _id: <ObjectId4>,
  id:  <ObjectId2>,
  gender:"Female",
  age:50
}
```

MORPH

```
{
  _id: <ObjectId2>,
  id: "022066_02M64",
  id_subject: "022066",
  gender: "M"
  age: 64
}
```

Figure 5.3: Implementation of MongoDB Service Collections: FGNET and MORPH Collections and the 1-to-many relationship with the DELETE collection.

### 5.2.4 Service Generator GUI

There is no GUI associated to the service evaluation platform. Only the command line interface (CLI) is available.

### 5.2.5 Plots

The implementation of the plots are done with the plotly library. Plotly is an interactive, open-source and browser-based graphing library for Python. Plotly provides online graphing, analytics and statistics tools. It requires a registration but has a free tier which has limited functionalities. The plots are categorised in the view layer where the credientials are invoked. The username and API key are required. Plots such as bar and box plot are implemented. For both graphs, the plot function receives two inputs: the data and the layout. The former is the set of pairs which could be for example the real age and the predicted age. The latter is the layout that of the plot that defines the type of plot, the labels, legends, titles, etc. An example of a plot can be found in `https://chart-studio.plotly.com/~felix.andabasabe/54/`

In Figure 5.4, a plotly graph example can be observed, there are several tabs from which by accessing the "Data" tab, the values of the generated graph can be seen. On the "Python & R" tab, it is possible to extract a dictionary format to export to either language and finally, the "Forking History" shows inputs from other collaborators.

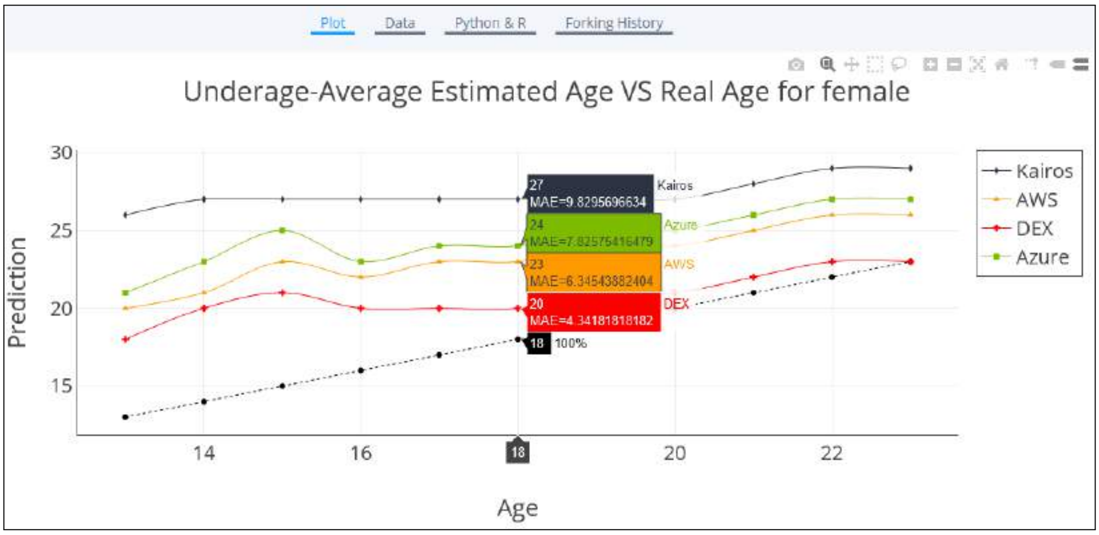


Figure 5.4: Plotly Graph

## 5.3 VisAGe - Dataset Framework

VisAGe is an image collection system designed for studying underage face subjects from 1 to 18 years old, in an unconstrained environment. The dataset consists of over 21,500 human-consensus-validated CC photographs of single faces collected from Flickr. Each face has been subject to a multiple-human verification process to ensure the reliability of age and gender annotations.

VisAGe is a web-based voting system aimed to label facial images with age and gender. The system has been built with a dataset of images that have been obtained from Flickr. The hybrid interaction with both humans and machines to tag data is an approach that can lessen the effort of users in order to contribute to research. The users are able to vote from a digital handheld device, laptop, computer, etc. Only a web browser is required. Once three votes are committed, the image becomes part of a positive voting dataset available for researchers. Moreover, different campaigns are created so the voting can be sorted by age, gender and type of faces such as single, multiple and others (No face). Finally, images are downloadable with the metadata corresponding to the Microsoft Azure Face API results. The results are presented in a JSON format and are attached to the files in a zip format. For more information about the user functionality, refer to the user manual in Appendix B.2.

### 5.3.1 Technical Overview

The structure of this system is divided in three main components, namely background scripts, data and web applications. The scripts are composed of a metadata collector, an image downloader and a facial feature detector. The data comprises both the VisAGe dataset (set of JPEG images) and a NoSQL MongoDB database. The web applications include the GUI and the RESTful web service. The following sections outline some technical details of the system.

The GUI was built on an open-source application Showoff [140], that was written in Python, utilising Flask. The front-end employs Bootstrap, jQuery and Swipebox. A crowd-sourcing approach was used to implement a voting scheme and establish the appropriate annotations of age and gender.

#### 5.3.1.1 Background Scripts

An MVC-architecture is used in the implementation of the background scripts. The model layer contains the definitions of photos, tags, Flickr API method, database, license, logs, etc. The controller layer contains the business application logic for the metadata collection and the image downloader. The main scripts are the following: **run_metadata.py, run_download.py** and **run_face_detector.py**. The full code is available at `https://bitbucket.org/4nd4/image-selector/`

The core functionality for the metadata collector, that integrates the Flickr API is the Method class. This class can be seen depicted in Listing 10. The method class is composed of three main parameters: the API key, the API secret, and the endpoint which refers to the URL of the server. The getter methods are related to the class attributes, the “get” method invokes the requests functionality from the Requests HTTP Library. The string passed to the requests.get method is composed by the Flickr  API method which can be “search”, the api key, the api secret, the arguments (which are passed as a parameter), and the format which is JSON. The response is recorded to the MongoDB collection.

The text argument for the “search” method to obtain the image with the desired age of a person was implemented with a regular expression which can be seen in Listing 9.

**Listing 9** Regular Expression for Age Text Search

```
^\d+((?i)st|nd|rd|th(?-i))+[[:blank:]]+(?i)\bbirthday\b(?-i)
```

The format accepts a number, the English ordinal suffix, a space and finally the word ”birthday”. Notice that including ordinals from different languages (French, German, Spanish, Chinese) and changing the birthday word to the corresponding word of the selected language would increase the number of images.

Flickr will return at most the first 4,000 results for any given search query. Therefore, two arguments are used to tweak the outcome: min_upload_date and max_upload_date. Both parameters are in the form of Unix timestamps and are regulated intelligently with a divide and conquer algorithm used to decrease the time complexity.

**Listing 10** Background Scripts Method Class

```
class Method:
    def __init__(self):
        self.method = None
        self.api_key = configuration.flickr_api_key
        self.api_secret = configuration.flickr_api_secret
        self.endpoint = 'https://api.flickr.com/services/rest/'
        self.child_class = self.__class__.__name__

    def set_method(self, method):
        child_class = str(self.__class__.__name__).lower()
        self.method = 'flickr.' + child_class + '.' + method

    def get_method(self):
        return self.method

    def get_api_key(self):
        return self.api_key

    def get_api_secret(self):
        return self.api_secret

    def get_endpoint(self):
        return self.endpoint

    def get(self, args=None):
        if args is None:
            args = ''

        return requests.get(
            self.get_endpoint() +
            "?method=" + self.get_method() +
            '&api_key=' + self.get_api_key() +
            '&api_secret=' + self.get_api_secret() +
            args + '&format=json' + '&nojsoncallback=1'
        )
```

The image downloader is also an MVC-based software design pattern. The python file *cn_image.py* belongs to the controller layer and has the processing method that selects the image metadata that was stored in the MongoDB collection and downloads from the recorded URL.

The feature detector was implemented with both `dlib` and the Microsoft Azure Face API. The former uses HOGs and a CNN approach. The latter is a cloud-based service that returns a structure such as the one in Listing 11 and is stored in the aggregate inside the image_created collection.

**Listing 11** Microsoft Azure Facial API Attribute Prediction Document Response.

```
{
    "faceId" : "f6679e39-60ac-ae03-82a78707f8b6",
    "faceRectangle" : { },
    "faceAttributes" : { },
    "faceLandmarks" : { }
}
```

#### 5.3.1.2 Data Records

The data records refer to both the non-relational database and the physical photographs. The non-relational database chosen is MongoDB; the preferred MongoDB GUI is Robo 3T (formerly known as Robomongo). This GUI is a visual tool aiding the management of Database MongoDB. It is a part of free open source software supporting all of three operating systems: Windows, Linux, Mac OS.

The VisAGe dataset contains over 21,800 human verified photos with age and gender labels; these annotated photos represent only 3.25% of the images that were downloaded from Flickr. The photos are single faces and have the respective age and gender label recorded in MongoDB.

The detailed NoSQL structure is divided into several collections. Due to privacy concerns, the GPS data has been stripped from each image which affects only 0.98% out of the entire VisAGe collection.

All the information concerning VisAGe, has been stored in a NoSQL server, divided into several collections. The collections with their descriptions can be seen depicted in Table 5.3.

The data files are compressed JPEG images. Each filename is named with an *ObjectId* which is a 12-byte unique value that consists of:

- a 4-byte value representing the seconds since the Unix epoch,
- a 5-byte random value, and
- a 3-byte counter, starting with a random value.

| Collection | Description |
|---|---|
| campaign | Refers to the combination of age and gender, and contains the positive and disagreed votes per campaign. |
| campaign_pending | Number of images left for voting per campaign per user. |
| data | Flickr metadata, original URL and annotated age. |
| deleted | Images deleted in a quality control. |
| image_created | Downloaded images and facial feature detection outputs are stored. |
| release | Information of the dataset release. |
| resolve | Resolution of disagreed votes. |
| toggle_resolution | temporal voting resolution records used to save the current state. |
| toggle_resolution_gender | temporal gender resolution records used to save the current state. |
| toggle_state | temporal state records of current votes. |
| vote | Image votes. |

Table 5.3: VisAGe Collections with the Respective Description. The image_created collection contains facial feature detection outputs.

The files are stored in a directory named after the age of the subject, which is an integer value; the JPEG files match with the id field of the *image_created* collection and a comma-separated values (CSV) file populated with metadata is created as per Table 5.4. The files are downloadable through the following frontend route in the Web Application:

```
/download/<age>/<gender>/<user_check>/<votes>
```

The *detail* field contains a JSON document with two main attributes, *dlib_service* and *azure_service*. The former consists of width and height values of the current image and the face bounding box detected by `dlib` with values such as width, top, height and left. The latter represents relevant face attributes obtained with Microsoft Azure Face API and can be seen represented in Table 4.5.

The python module used to compress the files is "zipfile"; The ZIP file format is a common archive and compression standard. This module provides tools to create, read, write, append, and list a ZIP file. When a dataset version is released, the relevant script is executed and the zip file is created with all the compressed images and attached metadata. Due to the amount of images compiled, the gathering process was programmed to run at midnight. This procedure would create each time a new release if the dataset was modified.

| Field | Description |
|---|---|
| id | Refers to the ObjectId. which is also the file name without the JPEG extension. |
| age | Age of the subject. |
| gender | Gender of the subject. |
| fli_license | Flickr CC License. |
| fli_username | Username that receives attribution for the usage of the image. |
| fli_uploaded_date | Value representing seconds since the Unix epoch for when the image was uploaded to Flickr. |
| fli_taken_date | Date the photograph was taken. |
| detail | Data captured using the facial feature detection tools. Both `dlib` and Microsoft Azure Face API, are structured in a JavaScript Object Notation (JSON) format. |

Table 5.4: VisAGe metadata format with the respective description

#### 5.3.1.3 Web Applications

Both the VisAGe dataset collection system and the Restful API are managed by the VisAGe GUI; specifically, in the controller layer on the file controllers.py. This file manages the frontend routes which are Python decorators (Refer to Section 4.4c) that Flask provides to assign URLs in the web app to function easily. Every web framework begins with the concept of serving content at a given URL.

Table 5.5 contains all the routes and descriptions for the VisAGe web application that are associated to an html template.

The final component of the system is a RESTful API which allows the system integration with other platforms. Consumers with access are able to retrieve the entire dataset URLs so that it can be downloaded from the original Flickr source with age and gender labels. The RESTful web service is accessible only for registered researchers through the VisAGe GUI application via the web service endpoint path: *`/gallery/api/v1.0/get_image_data/<int:face_rectangles>/<int:unique>/`*

The values *face_rectangles* and *unique* are replaced with either a *0* or a *1*. Setting *face_rectangles* to 1 will return the height, top, left and width values of the facial boundary box. Alternatively, when set to 0, data of the rectangle will not be retrievable but the response will be faster. Setting *unique* to 0 will return all the images; setting *unique* to 1 will only return a single image per Flickr user per age per gender. The latter feature is aimed to reduce bias by disallowing multiple images per user.

An example of a valid consumption would be the following endpoint path:

| route | description |
|---|---|
| ('/login', methods=['POST', 'GET']) | Login Page |
| ('/register', methods=['POST', 'GET']) | Register Page |
| ('/<campaign_id>/<filename>.html/<gender_edit>') | Main Image Page |
| ('/list/<campaign_id>.html') | List of Images per age |
| ('/') | Root |
| ('/campaign/<campaign_id>', methods=['GET', 'POST']) | Shows cover of album gallery |
| ('/statistics/<statistics_type>/') | Shows a page to choose the required statistics |
| ('/statistics_positives/<votes>/') | Lists statistics of positive voted files |
| ('/campaigns/') | for each campaign, get number of images left |
| ('/positive/<age>/<gender>/<votes>') | Lists positive voted files |
| ('/positive_resolved/<age>/<gender>/<votes>/') | Lists positive files that have been resolved due to disagreement |
| ('/disagreed/<age>/<gender>/') | Lists disagreed files that have both positive and negative votes |
| ('/gender_inaccurate/<age>/<gender>/') | List files where gender does not match |
| ('/campaign_description/<campaign_id>') | List of campaigns with the description |
| ('/search/') | Search functionality |

Table 5.5: VisAGe URL Routes

*/gallery/api/v1.0/get_image_data/0/0/*

## 5.4 Dlib Contour Artistic Approach

The DCA approach consists in alignment and landmark detection, both steps were implemented in python with the `dlib`, `cv2` and `numpy` libraries. The code can be found in the following link: `https://github.com/4ND4/DeepUAge`. The main files are *align.py* and *dca.py*.

### 5.4.1 Alignment

The alignment code can be seen in Listing 12. The alignment function *align_face_dlib* uses the `dlib` landmarks and has two input parameters: image and shape. The coordinates $(x, y)$ of the eyes are extracted and the indexes correspond to facial landmarks indexes defined in `dlib`. Once the coordinates are obtained, the angle between the eye centroids are computed with the *arctan2* function of `numpy`. Next, the median point is calculated between the two eyes in the input image. This value is stored in the *eyes_center* variable. Finally, the rotation matrix is obtained with the *getRotationMatrix2D* `cv2` function and the face is scaled with a warp affine transformation through the *warpAffine* `cv2` function.

**Listing 12** DCA - Align Code

```
from collections import OrderedDict
import cv2
import numpy as np

def align_face_dlib(image, shape):
    # extract the left and right eye (x, y)-coordinates

    (lStart, lEnd) = FACIAL_LANDMARKS_IDXS["left_eye"]
    (rStart, rEnd) = FACIAL_LANDMARKS_IDXS["right_eye"]
    leftEyePts = shape[lStart:lEnd]
    rightEyePts = shape[rStart:rEnd]

    left_eye_center = leftEyePts.mean(axis=0).astype("int")
    right_eye_center = rightEyePts.mean(axis=0).astype("int")

    # compute the angle between the eye centroids

    if left_eye_center is not None and
        right_eye_center is not None:

        dY = right_eye_center[1] - left_eye_center[1]
        dX = right_eye_center[0] - left_eye_center[0]
        angle = np.degrees(np.arctan2(dY, dX)) - 180

        # compute center (x, y)-coordinates
        # (i.e., the median point)
        # between the two eyes in the input image
        eyes_center = ((left_eye_center[0] +
                        right_eye_center[0]) // 2,
                       (left_eye_center[1] +
                        right_eye_center[1]) // 2)

        # grab the rotation matrix for rotating
        # and scaling the face
        M = cv2.getRotationMatrix2D(eyes_center, angle, 1.0)

        if image is not None:
            result = cv2.warpAffine(image, M, image.shape[1::-1],
                flags=cv2.INTER_LINEAR, borderValue=(0, 0, 0))

            return result
    else:
        print('error with eye center values')
```

### 5.4.2 Landmark Detection

The algorithm that emulates the face drawing methodology and predicts the hairline landmarks can be seen depicted in Algorithm 2.

**Algorithm 2** Dlib Contour Artistic Approach

```
procedure LANDMARK_DETECTION($image$)                ▷ Returns an array of vectors
    Obtain lowest point of the nose: $N = (x_{34}, y_{34})$
    Obtain closest points $[P_{left}, P_{right}] \perp N$
    Compute square side as per Equation 5.1
        $$sqr_{side} = \sqrt{(p_{left_x} - N_x)^2 + (p_{left_y} - N_y)^2} + \sqrt{(p_{right_x} - N_x)^2 + (p_{right_y} - N_y)^2} \quad (5.1)$$
    Build the square inscribed in a circle from vertex $v_1, v_2$, using $sqr_{side}$.
    Draw an icosagon from the centre of circle $C = (c_x, c_y)$.
    Return landmarks
```

The proposed method is based on the foundations of the Loomis face proportion approach. In Figure 5.5, the output of a digital sketch of the approach to detect hairline landmark points is depicted. The merge between the `dlib` face detector and the Algorithm 2 can be observed. Last, a cropped mask version of the face can be used to filter noise from facial images.

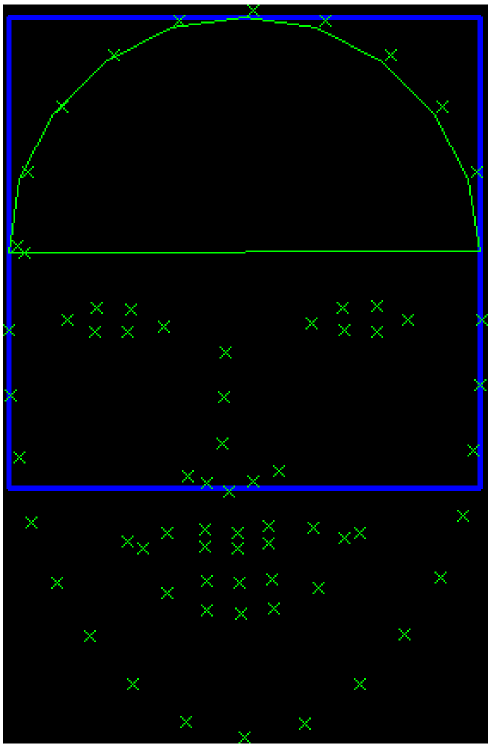


Figure 5.5: Contour Reconstruction from Dlib

## 5.5 Model Implementation

### 5.5.1 Early Approaches

A prototype that consumed the training and testing data from the Restful API mentioned in Section 5.3.1.3 was created and multiple architectures with/without data augmentation were tested. One of the earliest models developed was a simple stack of 4 convolution layers with a ReLU activation function followed by the max-pooling layers, a batch normalisation layer proposed by  [117] and dropout of 0.25 was added to help prevent overfitting[1]. This model used data augmentation (as outlined in Section 2.4.6.1) techniques and reached a MAE of 6.86 for the age group 1 to 25 years.

#### 5.5.1.1 Neural Network Architecture

An initial experiment to classify minors from adulthood led to a binary approach with a model similar to the back-propagation network proposed by LeCun et al. in 1989 [149] but with ReLU as the activation function, 3 stacked convolutional layers followed by maxpooling. Two binary models were created using both Caffe and Tensorflow. Training data consisted in a balanced dataset of 16,446 minors and adulthood. Evaluation data consisted in a balanced dataset of 1,440 minors and adulthood.

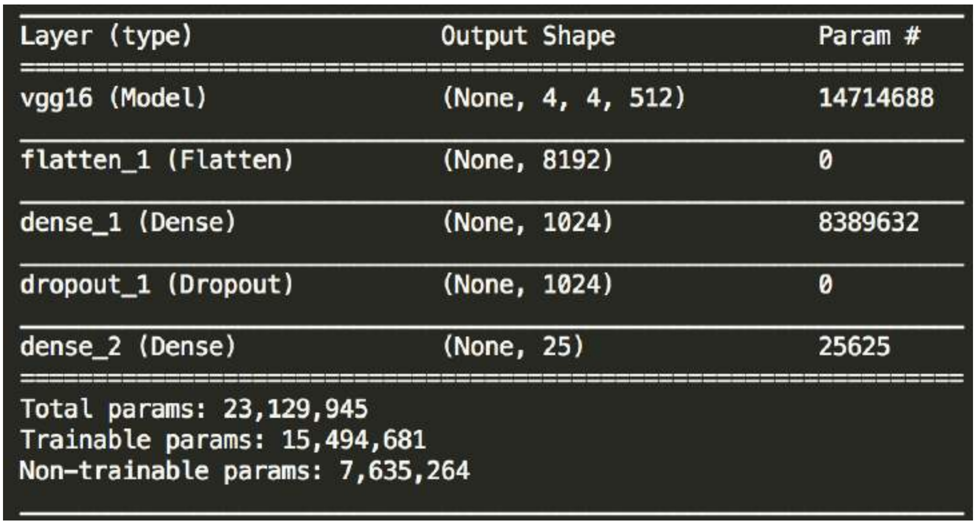
For the next experiment, a pre-trained VGG16 model with the ImageNet weights were used. The last 4 layers were modified to an output of 25 classes (ages) and the final activation function was set to softmax. The architecture is depicted in Figure 5.6. It can be seen that the total number of parameters are 23,129,945.

### 5.5.2 DS13K

The DS13K model was built with a VGG16 DCNN architecture. It was produced with the Caffe framework. Caffe is a DL framework made with expression, speed, and modularity in mind. It was developed by Berkeley AI Research (BAIR) and by a community of contributors. [129]. The source code can be found in the following URL: `https://github.com/4ND4/ds13k`

The DS13K architecture was designed in the *ageprototxt* protocol buffer definition file. An extract of the code can be seen depicted in Listing  13. The name of the architecture is "DS13K - VGG16", there are 4 *input_dim* values; the first corresponds to the batch

[1] Modelling error that occurs when a function is too closely fit to a limited set of data points



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
vgg16 (Model)                (None, 4, 4, 512)         14714688
_________________________________________________________________
flatten_1 (Flatten)          (None, 8192)              0
_________________________________________________________________
dense_1 (Dense)              (None, 1024)              8389632
_________________________________________________________________
dropout_1 (Dropout)          (None, 1024)              0
_________________________________________________________________
dense_2 (Dense)              (None, 25)                25625
=================================================================
Total params: 23,129,945
Trainable params: 15,494,681
Non-trainable params: 7,635,264
_________________________________________________________________
```

Figure 5.6: Architecture of Multi-Class Prototype Model

size, the second corresponds to the channels, and the last two values correspond to the image dimensions. The following are layers defined in a dictionary format.

**Listing 13** DS13K Architecture - Caffe Prototxt File for Age Estimation

```
name: "DS13K - VGG16"
input: "data"
input_dim: 1
input_dim: 3
input_dim: 224
input_dim: 224
layer {
  bottom: "data"
  top: "conv1_1"
  name: "conv1_1"
  type: "Convolution"
  convolution_param {
    num_output: 64
    pad: 1
    kernel_size: 3
  }
}
layer {
  bottom: "conv1_1"
  top: "conv1_1"
  name: "relu1_1"
  type: "ReLU"
}
layer {
  bottom: "conv1_1"
  top: "conv1_2"
  name: "conv1_2"
  type: "Convolution"
  convolution_param {
    num_output: 64
    pad: 1
    kernel_size: 3
  }
}
layer {
  bottom: "conv1_2"
  top: "conv1_2"
  name: "relu1_2"
  type: "ReLU"
}
layer {
  bottom: "conv1_2"
  top: "pool1"
  name: "pool1"
  type: "Pooling"
  pooling_param {
    pool: MAX
    kernel_size: 2
    stride: 2
  }
}
layer {
  bottom: "pool1"
  top: "conv2_1"
  name: "conv2_1"
  type: "Convolution"
  convolution_param {
    num_output: 128
    pad: 1
    kernel_size: 3
  }
}

.
.
.

layer {
  bottom: "fc7"
  top: "fc8-10ft"
  name: "fc8-10ft"
  type: "InnerProduct"
  inner_product_param {
    num_output: 5
  }
}
 layer {
  bottom: "fc8-10ft"
  top: "prob"
  name: "prob"
  type: "Softmax"
 }
```

The neural network architecture visualisation can be seen depicted in Figure 5.7. The online automatic tool used to create the *Caffe* graph is available at `http://yanglei.me/gen_proto/`

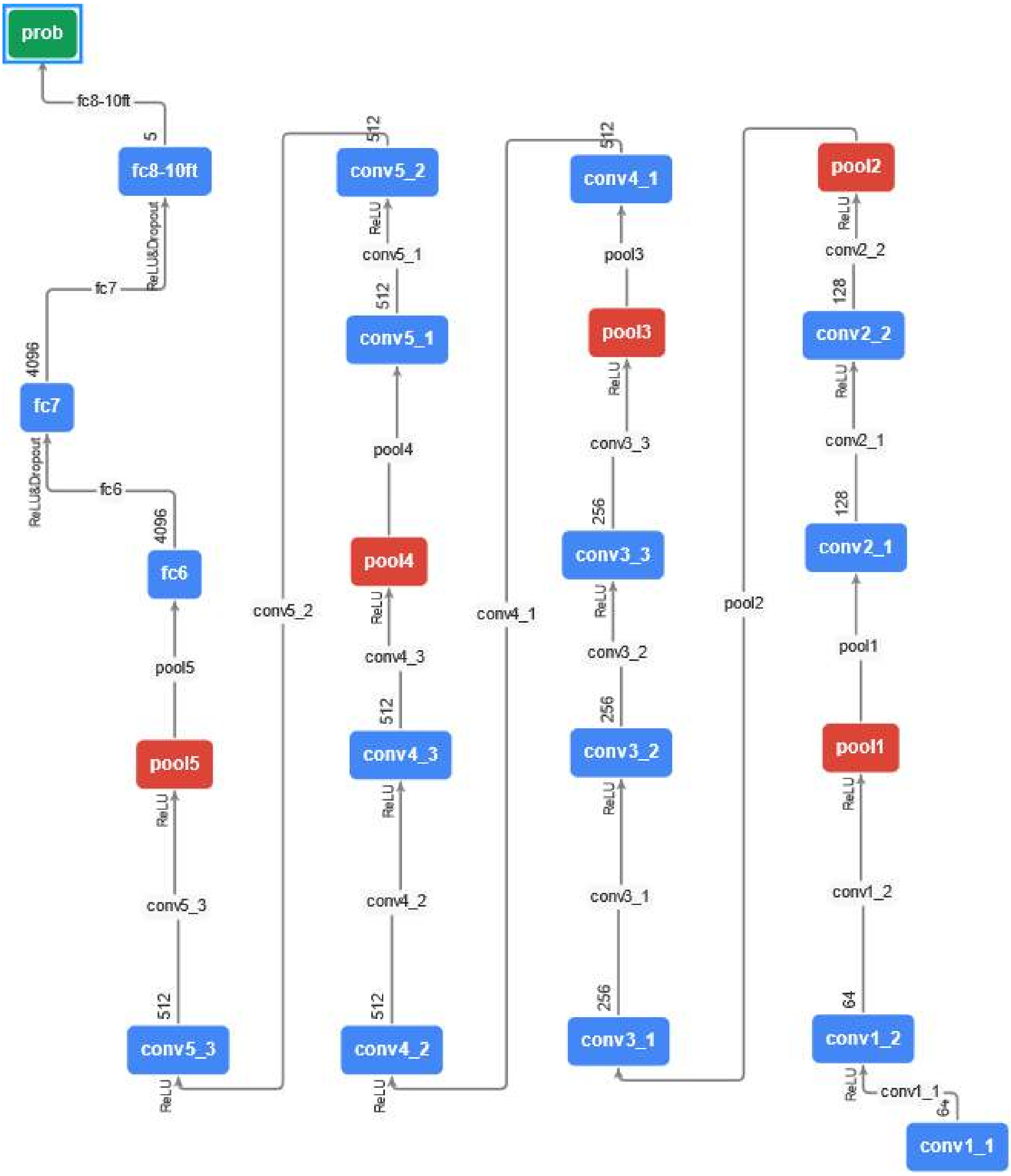


Figure 5.7: DS13K Caffe Graph

### 5.5.3 DeepUAge

The DeepUAge model was pre-processed by the DCA approach. The implementation of this approach was discussed in Section 5.4. Therefore, `dlib` was employed for facial detection. The images were converted to grey-scale and trained with the `keras` Model library.

The dataset split was performed with the script in Listing 14. The script iterates the data structure and separates the files randomly with the `random` library. Test and Training folders are managed separately. The files are stored in an array an then the random function is applied. The `shutil` package is employed to manage files. It can be seen that the rates configured are 0.8 and 0.2 for training and testing respectively.

**Listing 14** DeepUAge - Dataset Split

```
import os
import random
import shutil
import config

train_rate = 0.8
test_rate = 0.2

my_path = os.path.expanduser(config.mypath)
output_path = os.path.expanduser(config.output)
test_model_path = os.path.join(output_path, 'test')
train_model_path = os.path.join(output_path, 'train')

for _, dirs, _ in os.walk(my_path):
    for d in dirs:
        list_dir = os.listdir(os.path.join(my_path, d))

        dataset_length = len(list_dir)

        train_length = int(dataset_length * train_rate)

        arr = [x for x in range(0, len(list_dir))]

        random.shuffle(arr)

        print('creating training set...')

        normalize_folder_index = 1

        for i in arr[:train_length]:
            .
            .
            .
            shutil.copy2(source_path, destination_path)

        print('creating testing set')

        for i in arr[train_length:dataset_length]:
            .
            .
            .
            shutil.copy2(source_path, destination_path)
```

The outcome of the script produces a structure as per Figure 5.8. The files are stored in their correspondent age folder from 1 to 20, and testing and validation has the similar folder structure as training (omitted for presentation).

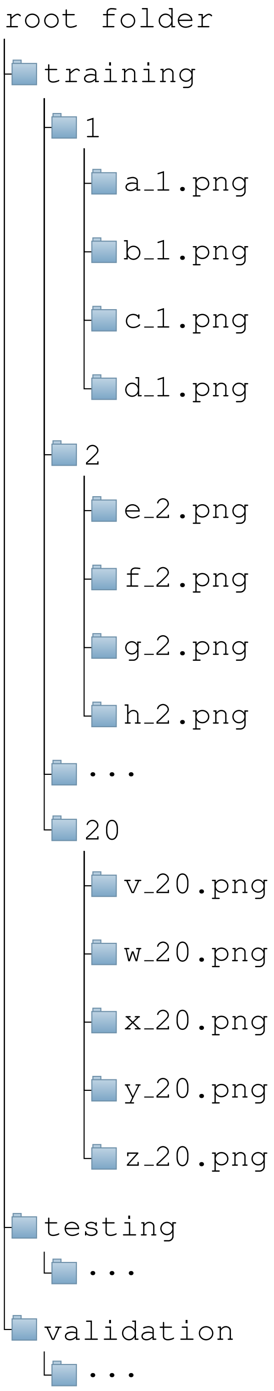


Figure 5.8: DeepUAge Folder Structure

Data augmentation techniques such as flip, rotation, zoom, distortion, colour, contrast, brightness and random erasing were used. These techniques were implemented with the `Augmented` python library creating new images "on the fly". The function used for the augmentation transformations can be seen depicted in Listing 15.

**Listing 15** DeepUAge - Dataset Augmentation

```python
def get_transform_func():
    p = Augmentor.Pipeline()
    p.flip_left_right(probability=0.5)
    p.rotate(
        probability=1,
        max_left_rotation=5,
        max_right_rotation=5
        )
    p.zoom_random(probability=0.5, percentage_area=0.95)
    p.random_distortion(
        probability=0.5,
        grid_width=2,
        grid_height=2,
        magnitude=8
        )
    p.random_color(
        probability=1,
        min_factor=0.8, max_factor=1.2
        )
    p.random_contrast(
        probability=1,
        min_factor=0.8, max_factor=1.2
        )
    p.random_brightness(
        probability=1,
        min_factor=0.8, max_factor=1.2
        )
    p.random_erasing(probability=0.5, rectangle_area=0.2)

    def transform_image(image):
        image = [Image.fromarray(image)]
        for operation in p.operations:
            r = round(random.uniform(0, 1), 1)
            if r <= operation.probability:
                image = operation.perform_operation(image)
        return image[0]
    return transform_image
```

The age estimation approach was addressed as a classification problem and the model was trained with the ResNet50 module from `keras`. Transfer learning was used and the base model was set to not include the top layer in which case the FC output layers of the model used to make predictions is not loaded, allowing a new output layer to be added and trained. This last layer is of size 20. The weights of ImageNet are loaded and the input shape is of (224, 224, 3) where 224 x 224 refers to the image size and 3 refers to the number of channels. The base model creation can be seen depicted in Listing 16.

Finally, the hyper-parameters were managed mainly with `keras` libraries such as *keras.callbacks* and *keras.optimizers*. Early stopping, learning rate scheduler and model checkpoint was accomplished with the former library. The optimiser (SGD) was implemented with the latter.

**Listing 16** Keras ResNet50 Model Creation

```
from keras.applications import ResNet50

base_model = ResNet50(include_top=False, weights='imagenet',
 ↪  input_shape=(image_size, image_size, 3), pooling="avg")
```

### 5.5.4 Vec2UAge

#### 5.5.4.1 Dataset Curation

A file format normalisation was required. The format of the Selfie-FV files are as follows: **[artist_name]_[age]_[series].jpg**. The Series is a number that indicates the sequence of the image. This number is not important for the file format normalisation. Therefore, it was set to “X”. The “artist_name” was set to a unique identifier. A script was created to normalise the visage files. This script can be seen in Listing 17.

It can be seen that the `os`, `shutil` and `uuid` libraries are employed. The first library mainly manages paths, the second library manages the copy process and the last library creates a universally unique identifier (UUID) which is a 128-bit number. The directory is expected in a structure seen in Figure 5.9. The script iterates each directory, for each directory each file is formatted to the new standard. The *string.format* function is used. Line 23 displays how the new name is created and line 25 shows how the file is copied to a new single directory.

**Listing 17** File Normalisation Script

```python
import os
import shutil
import uuid

for d in range(1, 19):
        directory_visage = os.listdir(
            os.path.join(input_folder, str(d))
        )

        for f in directory_visage:
            if f.startswith('.'):
                continue

            file_name, file_extension = os.path.splitext(f)

            unique_filename = str(uuid.uuid4().hex)

            new_name = '{}_{}_x{}'.format(unique_filename, d, file_extension)

            if not os.path.exists(output_folder):
                os.mkdir(output_folder)

            new_path = os.path.join(output_folder, new_name)

            shutil.copy2(os.path.join(input_folder, str(d), f), new_path)

            print(f, 'processed')
```

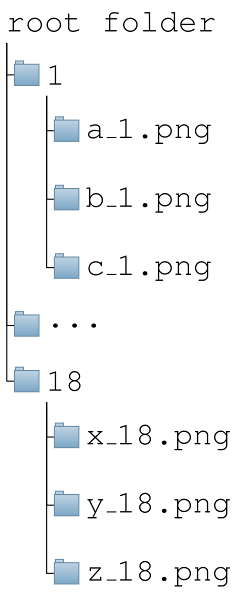


Figure 5.9: VisAGe Image Structure before Merge.

Once the files were normalised and copied to a single folder, the next step was to merge the VisAGe (outlined in Section 4.3.2) and Selfie-FV (outlined in Section 3.3.4.1) datasets into the aforementioned single folder. This process may be done manually or with the `shutil` library.

#### 5.5.4.2 Dataset Pre-Processing

The Vec2UAge model was pre-processed by the `dlib` library. When no face is detected, the CNN version of `dlib` is executed. This approach lessens the processing time and increases the face recognition hits. The code of both types of `dlib` approaches can be seen in Listing 18. The second approach requires the weights file *mmod_human_face_detector.dat*.

**Listing 18** Dlib Facial Detection Types

```python
from imutils import face_utils
import dlib

def dlib_detector(img):
    if img is not None:
        detector = dlib.get_frontal_face_detector()
        rects = detector(img, 1)

        # loop over the face detections
        for (i, rect) in enumerate(rects):
            return face_utils.rect_to_bb(rect)


def dlib_cnn_detector(img):
    WEIGHTS = 'preprocessing/mmod_human_face_detector.dat'

    cnn_face_detector =
     ↪  dlib.cnn_face_detection_model_v1(WEIGHTS)

    dets = cnn_face_detector(img, 1)

    # loop over the face detections

    for i, d in enumerate(dets):
        x = d.rect.left()
        y = d.rect.top()
        w = d.rect.right() - x
        h = d.rect.bottom() - y

        return x, y, w, h
```

After the faces were detected, the face vectors were calculated and stored in a numpy array with functions in Listing 19. The *get_vectors_np* function obtains the vectors in a numpy array format. The input parameters are the face images input path and the image size. The *faces_to_vectors* function requires the path of the model. The pre-trained FaceNet model is available in the following URL: `https://github.com/davidsandberg/facenet`. Given a folder and a model, the function loads images and performs a forward pass to get a vector for each face. This function uses the `tensorflow` library and returns the embedded array.

**Listing 19** FaceNet Vector Extraction

```
def faces_to_vectors(inpath, modelpath, imgsize):

    tf.disable_v2_behavior()
    with tf.Graph().as_default():
        with tf.compat.v1.Session() as sess:
            load_model(modelpath)
            image_paths = get_image_paths(inpath)
            images_placeholder =
            tf.get_default_graph().get_tensor_by_
            name("input:0")

            embeddings =
            tf.get_default_graph().get_tensor_by_
            name("embeddings:0")
            phase_train_placeholder =
             ↪  tf.get_default_graph().get_tensor_by_
            name("phase_train:0")

            images = load_data(image_paths=image_paths,
             ↪  do_random_crop=False, do_random_flip=False,
                               image_size=imgsize,
                                ↪  do_prewhiten=True)
            feed_dict = {images_placeholder: images,
             ↪  phase_train_placeholder: False}

            emb_array = sess.run(embeddings,
             ↪  feed_dict=feed_dict)

            return emb_array

def get_vectors_np(input_path, image_size):

    mdlpath = 'models/facenet/20180402-114759.pb'

    vectors = faces_to_vectors(
        inpath=input_path, modelpath=mdlpath,
         ↪  imgsize=image_size
    )

    return vectors
```

#### 5.5.4.3 Dataset Split

The dataset split was managed by a stratified split shuffle function in the main training file. The code can be seen in Listing 20. The function is part of the *sklearn.model_selection* library and requires the number of splits, the test size which represents the split ratio and a random seed.

**Listing 20** Vec2UAge - Dataset Split

```
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.5,
 ↪  random_state=random_seed)
```

#### 5.5.4.4 Dataset Augmentation

Data augmentation is managed by the *Augmentor* ML python library [26]. *Augmentor* aids the image augmentation and artificial generation of data for machine learning use cases. It uses a stochastic approach using building blocks that enable operations to be pieced together in a pipeline.

The cosine similarity between the original image and the augmented image was computed with the Equation 3.1. As a reference, the vectors from Figure 4.18 were calculated and the cosine similarity was analysed. The results can be seen depicted in Table 5.6. If the cosine similarity is close to 1, the augmented facial vector has not suffered much changes and would resemble the original image. Therefore, a more aggressive augmentation technique would be required. According to the results in the table, it can be observed that FaceNet is robust against occlusions. Therefore, the random erasure data augmentation technique should be replaced by another augmentation method.

After the analysis, image augmentation was selected accordingly to a defined threshold of 0.6 – meaning that the euclidean distance between the facial vectors predicted from the images and augmentations were slightly far from each other. However, the augmented dataset was only used for the training set as discussed previously. These techniques were performed in an offline manner. Thus, creating physical images saved to the local disk.

| **Augmentation** | **Cosine Similarity** | **Settings** |
|---|---|---|
| Flip | 0.8599 | Horizontal |
| Brightness | 0.6845 | Factor: 2 |
| Rotation | 0.6656 | Angle: 25 |
| Random Zoom | 0.7856 | Factor: 2 |
| Random Distortion | 0.8728 | Grid width: 10<br>Grid height: 10<br>Magnitude: 8 |
| Random Colour | 0.5609 | Factor: 2 |
| Random Contrast | 0.3341 | Factor: 5 |
| Random Erasure | 0.9837 | Rectangle : 0.2 |

Table 5.6: Cosine Similarity between Original Image and Augmented Images Using the *Augmentor* library

#### 5.5.4.5 Network Architecture

The DL framework used to train the models was managed by the `pytorch_lightning` (pl) library. This is an open source Python library that provides a high-level interface for PyTorch. A *Vec2UAgeSystem* class that extends the `pl.LightningModule` class is initialised with hyperparameters such as LR, batch size and optimiser choice. The full code is available on github in the following URL: `https://github.com/4ND4/Vec2UAge`.

#### 5.5.4.6 Hyper-Parameters

The experiments were managed with Neptune. Pytorch_lightning and Neptune can be integrated through the *pytorch_lightning.loggers.NeptuneLogger* class. The Neptune logger can be used in the online mode or offline (silent) mode. To log experiment data in online mode, NeptuneLogger requires an API key. In offline mode, the logger does not connect to Neptune. The Neptune connection can be seen depicted in Listing 21.

**Listing 21** Neptune Configuration

```
tags = [opt_choice, unique_tag]

neptune_logger = NeptuneLogger(
            api_key=config.API_KEY,
            project_name=project_name,
            experiment_name=experiment_name,
            close_after_fit=False,
            params={
                "epochs": epochs,
                "patience": patience,
                "number_layers": nb_layers,
                "batch_size": config.BATCH_SIZE,
                "optimizer": opt_choice,
                "learning_rate": learning_rate,
                "json_path": [train_json, test_json],
                "early_stop": early_stop,
                "random_seed": random_seed
            },
            tags=tags

        )
```

The API key is obtained when registering to Neptune. A tier is available for researchers at no cost. The project name refers to descriptive name for the project. The experiment is more specific. For example, The project name is **Vec2Uage** and the experiment is "A1". The parameter *close_after_fit* regulates if the experiment would be closed after training. If False, the experiment will not be closed after training and additional metrics, images or artifacts can be logged. The *params* parameter is a dictionary that contains information that would be logged directly to the experiment. Keys such as epochs, patience, number of layers, batch size, optimiser, LR, JSON path, early stop and random seed are logged into the Neptune experiment. Tags help identify the experiment by the optimiser choice, unique tags, etc.

#### 5.5.4.7 Improving Traditional Hash Sets with Facial Embedding Vectors

The Vec2UAge model approach used data augmentation techniques and the cosine similarity was computed between the original image and the augmented image to detect if the original image had significant changes. This function inspired the design of an improved functionality of the traditional hash set comparisons. Traditional Hash sets were discussed in Section 2.2.6.1. The hash value approach is highly static and limited. A database of facial features could be used to identify victims and missing persons. The facial embedding image recognition approach calculates the vector of a given suspect/victim and then initiates a comparison given a function that is mapped to the vector database. This method is able to compare if there is a match with-in the database. The algorithm can be seen in Algorithm 3.

**Algorithm 3** Facial Embedding Image Recognition Approach - FaceNet

```
procedure FACE_COMPARE($image$)                                   ▷ Returns a boolean
    Detect and crop $face$ in $image$
    Resize $face$ to 224 x 224 pixels
    Extract feature vector $p$ from $face$
    Compute euclidean distance between feature vectors (Equation 2.1):
    while $true$ do                                ▷ Iterate until end of feature database
        if $d(p, q) < threshold$ then
            Return $true$
```

CHAPTER

SIX

# RESULTS

## 6.1 Facial Age Estimation Model Evaluation Platform

This section outlines the results from the evaluation of the current online and offline age estimating options. Firstly, the results are presented across the entire age range of the dataset (0-77 years old) in Section 6.1.2. Next, Section 6.1.3 presents the results of subdividing the dataset by gender. Finally, in Section 6.1.4 the performance of the four prediction systems within different age ranges are compared (Amazon Rekognition hereafter called AWS, Azure, Kairos and DEX). It is important to consider that this evaluation answers the question of what the performance is. Not the "why" or "how" the performance was obtained. This is due to the online cloud-based age estimation services being "blackboxes" for which all that is exposed is the performance.

### 6.1.1 Dataset Experiment Summary

The dataset used in this section is the one produced by the dataset generator discussed in Section 4.2. The several dataset sources were a combination of the datasets presented in Section 3.5 (FG-NET, FERET, MEDS, IMDB, WIKI, YFCC100M, etc.). As mentioned in the evaluation performance methodology in Section 4.2.5, a random equally distributed dataset (for age classes) was required. Initially, a collection of images from the whole age range of 0 to 100 was considered. However, due to lack of images (visible in Figure 4.3), an age range from 0 to 77 was studied. Due to a scarcity of images of children between the ages of 0 and 14 and the applicability of this age range to CSEM investigation, additional images were manually collected. CC licensed pictures with accurate age and gender were gathered from Flickr. Photos were manually labelled with the age and gender of the individual based on the descriptions, title of the photo, the respective tags or any visual clues. This procedure resembled how the VisAGe dataset was collected (discussed in Section 4.3.2). The process ultimately ensured that 65 images per age per gender could be included. Resulting in a total dataset size of 10,140 images.

Each image was passed through each of the four systems mentioned in Section 6.1, and the results recorded. These were then evaluated under three influencing factors. Initially, the four systems were compared across the entire age range by analysing the error rates that each exhibited, i.e. the difference between the predicted age of each subject and their actual age. Next the dataset was divided by gender to investigate whether this had any effect on the accuracy of the age predictions. Finally, in the third test the dataset was subdivided into a number of age ranges, i.e. 0-9, 10-19, 20-29, etc. The goal here was to find whether certain systems performed better in different age ranges, or whether one system could be said to be the most accurate over the entire dataset.

### 6.1.2 Entire Age Range Estimation

The first analysis that was conducted was to measure the accuracy associated with each of the four systems across the entire dataset (outlined in Section 6.1.1), with a view to discovering which services are most effective. Firstly, the MAE was calculated for each system across all the subjects. The results of this are shown in Table 6.1.

Table 6.1: Mean Absolute Error per Service.

| **Service** | **MAE** (ages) |
|---|---|
| Kairos | 11.236 |
| AWS | 9.286 |
| DEX | 8.079 |
| Azure | 7.614 |

Using this metric, Microsoft Azure was found to achieve the best results, with the lowest MAE. This can be interpreted to mean that across the entire range of all subjects, the average difference between the predicted age and the actual age was 7.614 years. The error rates for the other services were higher, with DEX achieving better overall results compared to AWS, which in turn outperformed Kairos. While this is a useful finding in itself, a more in-depth view is required to examine the characteristics of each service further.

Figure 6.1 illustrates the performance of the four systems more clearly across the dataset. The X-axis indicates the actual age of the subjects. A line is plotted for each service, which indicates the average age it predicted for subjects in each age class. Each point therefore represents the average predicted age for 65 subjects that have the same age. The dotted line is used to indicate where correct predictions should lie.

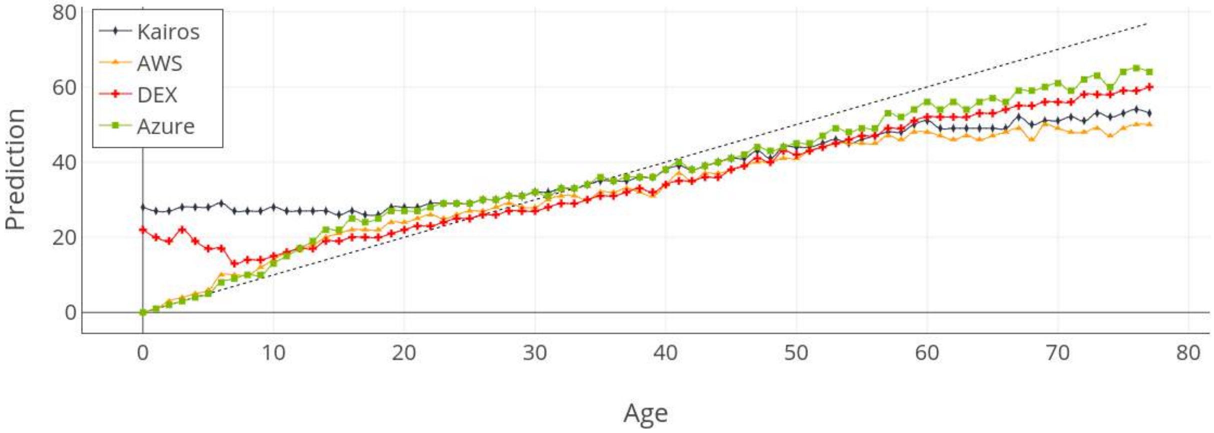


Figure 6.1: Average Estimated Age Compared with Actual Age across Entire Dataset.

A number of interesting observations can be made from this figure. Both DEX and Kairos have a tendency to substantially over-estimate the age of young children. In the case of Kairos, this over-estimation continues well into the late teenage years, before its predictions become closer to those of its competitors after this point.

In all cases, a tendency to underestimate the age of subjects begins to emerge from approximately the age of 40, though this is more pronounced for some services. Kairos, in addition to being the least accurate at early age ranges, also has the second-highest error rate for older subjects, behind only AWS. DEX, from having a high error rate for young children, becomes the second most accurate (behind Azure) at later ages. Its tendency to underestimate ages becomes apparent earliest, from the late-20s onward. The line for AWS is very close to the true age line in the early stages, but exhibits the highest level of underestimation for the later ages.

When bearing digital forensic use-cases in mind, it is worthwhile focusing on the late teenage years in particular, around the boundaries where people cease to be minors in various jurisdictions. The very accurate performance of AWS and Azure begins to diverge from the correct prediction line around the age of 10. DEX, which performs poorly on young children, is closest to this ideal line in the mid-to-late teenage years and continues into the early 20s. A further examination of the relative performance of the systems in various age ranges is contained in Section 6.1.4.

### 6.1.3 Influence of Gender on Estimation

To further explore the characteristics of the four services, the dataset mentioned earlier was classified by gender, and a similar analysis to the previous section was conducted. The overall MAE for each service is shown in Table 6.2. The main interesting insight that can be gained from this table is that uniformly, all four services exhibit a higher rate of error for female subjects than for males. Of these, Kairos is the only one for which the difference in error rates by gender is less than 1 year on average. The difference is most pronounced for AWS, for which the error rate for male subjects is only marginally greater than for DEX, but whose predictions for female subjects are more comparable to Kairos. The relative ranking of the four systems remains the same for both genders.

Table 6.2: Mean Absolute Error per Gender per Service.

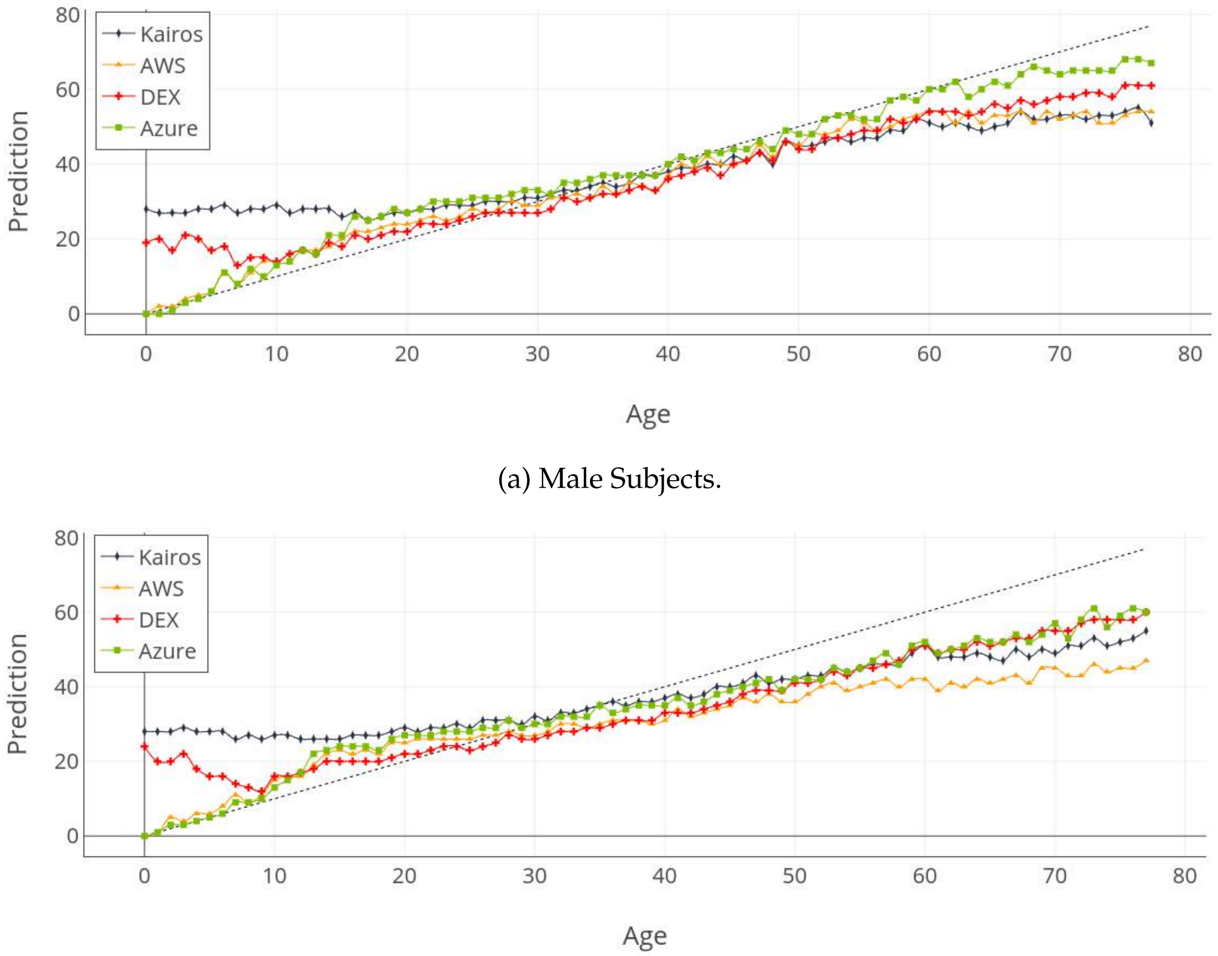
| Service | Male | Female |
|---|---|---|
| Kairos | 10.6838 | 11.7960 |
| AWS | 7.2192 | 11.4057 |
| DEX | 7.1975 | 8.9613 |
| Azure | 6.4205 | 8.8092 |

As with the previous section, a more in-depth view is required beyond the overall error rates. Figure 6.2 is constructed in the same way as for Figure 6.1, with the exception that it shows separate line graphs for each gender. Figure 6.2a refers only to the results for male subjects, and Figure 6.2b relates only female subjects.

The overall patterns observed in the previous section are generally apparent in both graphs: Kairos and DEX overestimate at young ages and all four services tend to underestimate the age of older subjects. However, the rate at which the latter effect occurs is far more pronounced for female subjects. By the age of 36, all four services underestimate on average, and this gap becomes more pronounced with increasing age. In contrast, Azure in particular remains much closer to the ideal line for male subjects.

Figure 6.3 illustrates this data by displaying the MAE rate for each of the systems at each age. Error rates generally increase towards older ages, with this being more pronounced for female subjects, due to the age underestimation common to all services. As previously observed, Kairos and DEX exhibit relatively high error rates for young subjects. However, it is notable that although Kairos is clearly the least accurate for young subjects, it achieves better error rates than the other systems towards the middle of the ages evaluated. This is the focus of the following section, where this data is viewed within a range of age brackets.

A local peak is observed in the teenage years, before error rates decline into the 20s and



(a) Male Subjects.

(b) Female Subjects.

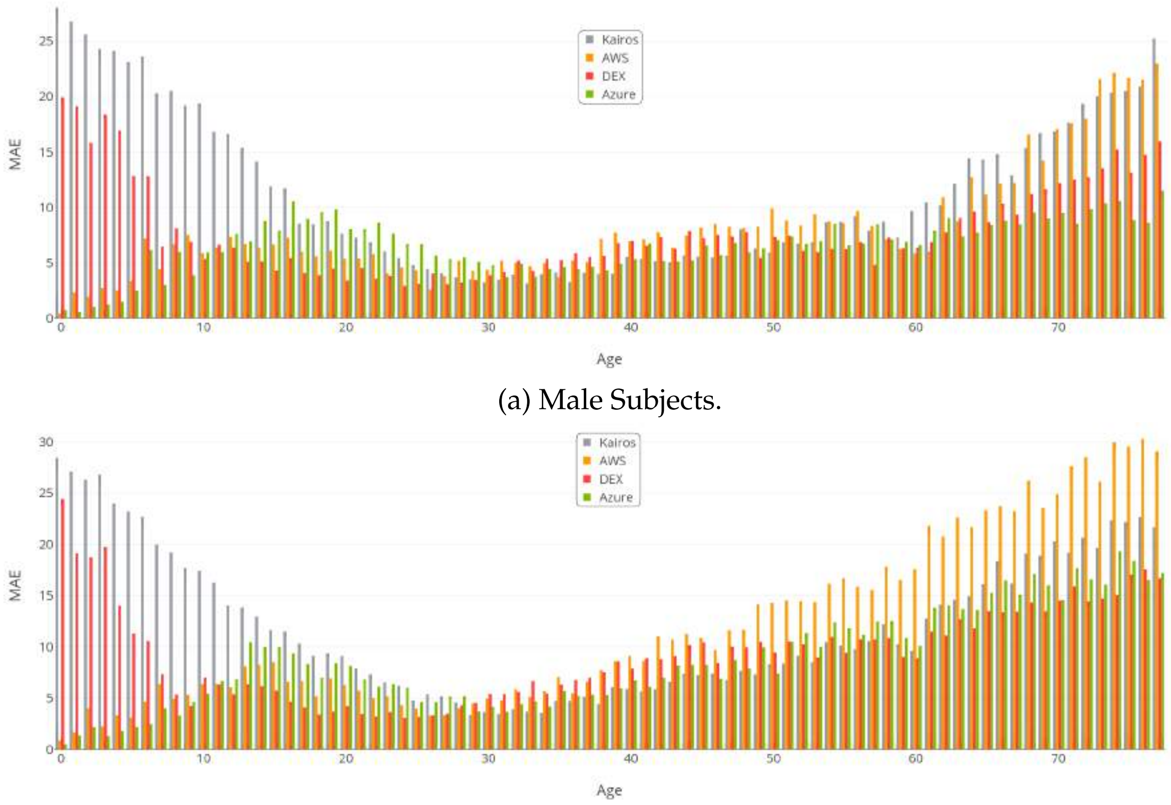
Figure 6.2: Average Estimated Age Compared with Actual Age

30s. This suggests that more focus is required on this area in the future, especially due to the use cases that require accuracy within this borderline adulthood age range.

### 6.1.4 Age Range Analysis

Previous observations of the data indicate that although the Azure performed better than the other four on average, performance was affected not only by gender but also according to age. This motivated a deeper analysis of the relative performance of the four systems across different age ranges. For this, the aforementioned dataset was subdivided into 10-year age ranges (0-9, 10-19, 20-29, etc.). The only exception was the final range, which was from 70 to 77 due to the lack of available older subjects in constituent datasets.

Figure 6.4 was generated to provide insights into the data. In this figure, a box is plot-



(a) Male Subjects.

(b) Female Subjects.

Figure 6.3: Mean Average Error Rate

ted for each service within each age range. For each box, the data used was the average predicted age for each actual age. The boxes show the mean, median, interquartile range, with the whiskers representing the maximum and minimum values within 1.5 times of the interquartile range. Outliers are shown as individual points where relevant.

The primary object of this study is to ascertain which system(s) offer the most accurate performance for the age prediction task. Although Azure has been shown to the lowest overall error, this figure indicates that this does not reflect an overall superior performance across all age ranges. The best-performing system, on average, in each age range is summarised in Table 6.3. Azure is the most precise only for the youngest and oldest age ranges. A somewhat surprising result here is that although Kairos has the highest overall error rate, it is the most precise on average in the 30-39 and 40-49 age ranges. The high rate of error for DEX for the youngest children is much reduced by the teenage years, and this system has the lowest error rates for the 10-19 and 20-29 age ranges. Indeed, it is again the most precise for 50-59 and 60-69. In total, DEX has the lowest error rate for four of the age ranges, with Azure and Kairos having the low-

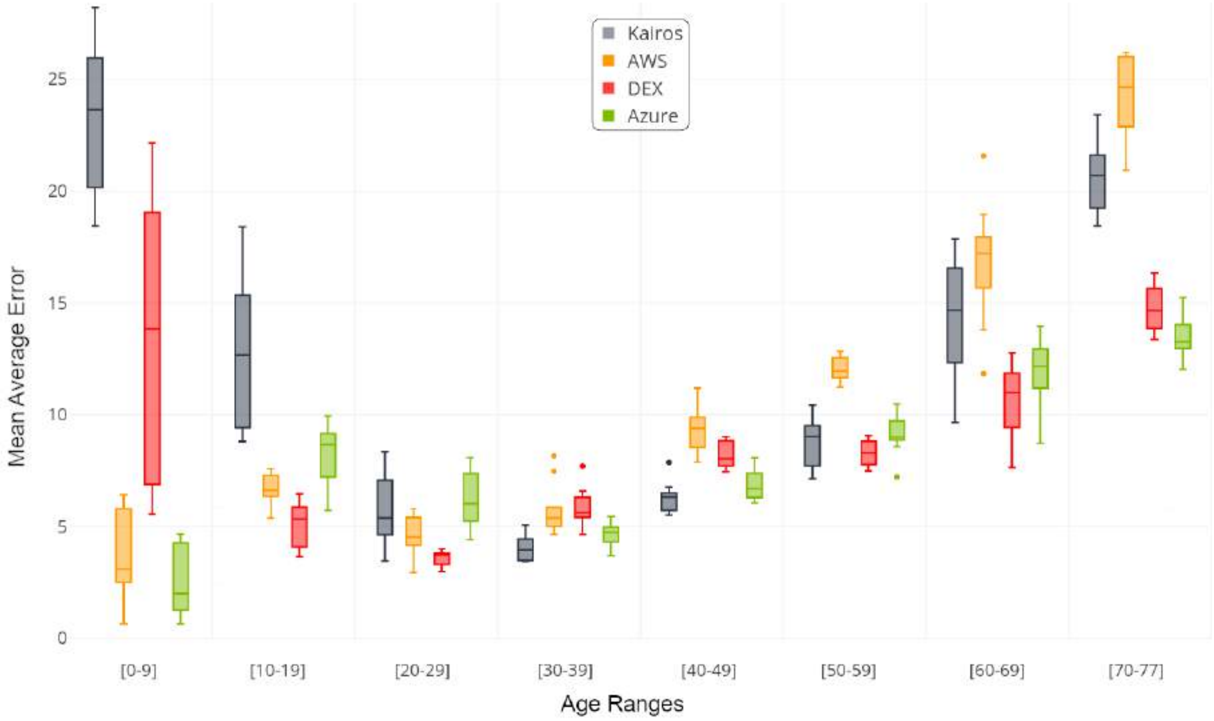


Figure 6.4: Mean Average Error Rate for Each System in Different Age Ranges.

est error for two range apiece. AWS is not the most accurate in any of the ranges, but is the second-best average performer for all ranges up to the age of 29.

Table 6.3: Performance per Age Range.

| Age Range | Best Performer |
|---|---|
| 0-9 | Azure |
| 10-19 | DEX |
| 20-29 | DEX |
| 30-39 | Kairos |
| 40-49 | Kairos |
| 50-59 | DEX |
| 60-69 | DEX |
| 70-77 | Azure |

## 6.2 Underage Age Estimation Improvement

Three new set of experiments were conducted and the MAE was calculated. The results of which are presented in the subsections that follows. The first experiment, discussed in Section 6.2.2 focused on the wider age range from 0 to 25 years old, to evaluate and compare the four individual services: How-Old.net, AWS, DEX, and Azure. The

Kairos Service was replaced by How-Old.net due to low performance discovered in Section 6.1.2. The second experiment involves the evaluation of DS13K. The model is evaluated in Section 6.2.3. The final experiment introduces ensemble ML techniques to establish whether these will be useful tools to improve upon the performance of the four systems. This is presented in Section 6.2.4.

### 6.2.1 Dataset

The dataset used in this section is the DS13K dataset with a size of 12,792 images (categorised as a small dataset per Section 3.3) and described in Section 4.5.1.1.

### 6.2.2 Underage Range Baseline Estimation

The evaluation for the first experiment focused on samples from 0 to 25. The results of the evaluation are shown in Figure 6.5, with the average predicted age for each service plotted against the subjects' actual ages. The MAE for each service can be seen in Figure 6.6 and the average MAE for underage subjects is presented in Table 6.4.

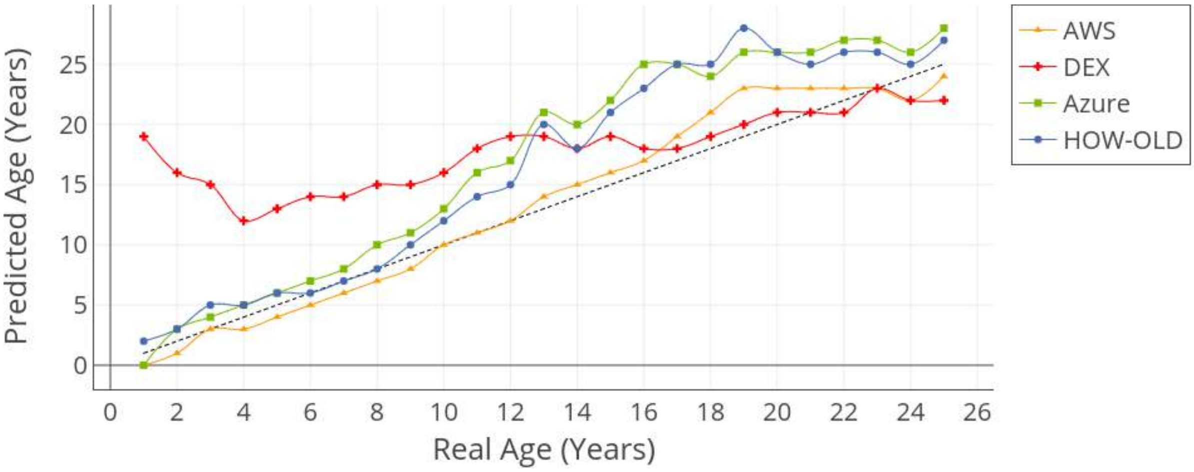


Figure 6.5: Average Estimated Age from Each Service Compared with Actual Age.

From these figures, it can be seen that AWS performs best overall. Although it has a slight tendency towards underestimation up to the age of 12, it maintains its accuracy in older age groups better than Azure and How-Old.net, whose predictions gradually deviate away from the real age between the ages of 10 and 22. These three services show similar accuracy for the youngest subjects below the age of 12.

In contrast, DEXs pretrained model fails to accurately classify the younger samples. However, from 17 to 21 years old (in the crucial underage/adulthood boundary zone), it has a better performance than the rest of the models. This pattern is likely due to

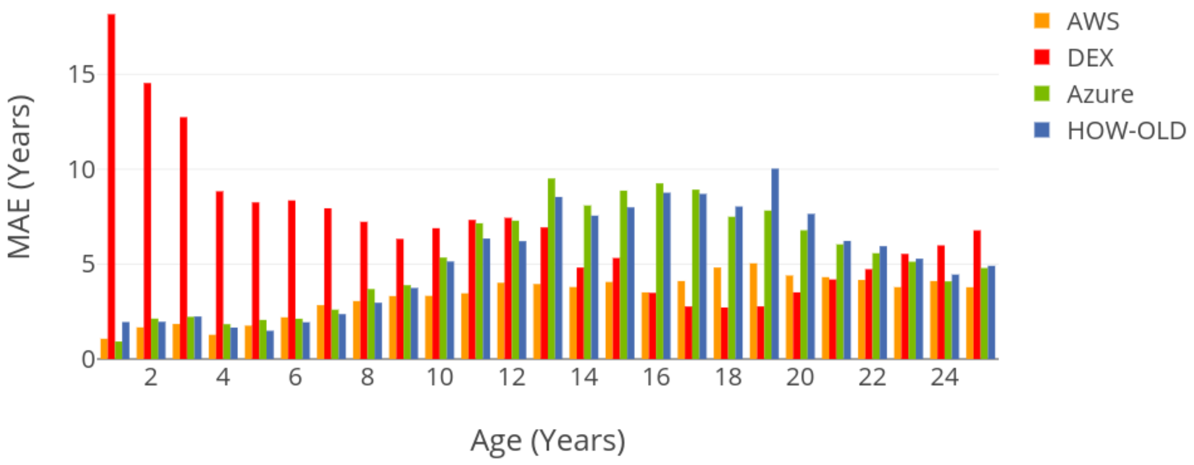


Figure 6.6: Mean Absolute Error per Age by Service.

a lack of sufficient sample images used to train the Deep Expectation model for very young subjects, and is the primary reason why DEXs overall MAE is higher than the others.

In terms of overall MAE for underage subjects, the AWS biometric detector service performed better than the rest of the services with a MAE of 3.347 as shown in Table 6.4. Although the output of the prediction accomplished by AWS was classified with a high and low range, we found that the closest value to the real age would be the lowest value. AWSs superiority is unrivalled across the majority of age ranges, in fact it is between the best two performers for each age. It is also observed that only DEX and AWS underestimated the subjects' ages at any point, while the remaining services overestimated the values almost throughout the entire age range.

| **Service** | **MAE** |
|---|---|
| AWS | 3.349 |
| How-Old.net | 5.281 |
| Microsoft Azure | 5.347 |
| DEX | 6.936 |

Table 6.4: Mean Absolute Error for Underage Images per Service.

### 6.2.3 DS13K Evaluation and Comparison

The accuracy per age group as well as the average accuracy per service is depicted in Table 6.5, where the best-performing figure for each age range is illustrated in bold. Our developed DS13K model has the best average performance followed closely by AWS. However, looking at the aforementioned table, the winner in several age groups is AWS only dropping for the case of 16-17 year-old's. In the key [16-17] age range, the accuracy of DS13K was substantially higher than the other services, with 68% of subjects in this range being successfully classified. This improvement is still low performance when considering digital forensic use cases and a threshold for an acceptable accuracy rate should be established. The second-highest accuracy for this range was AWS with 15%. As illustrated previously in Figure 6.5, all the other services tend to overestimate age for subjects in this range, which would lead to underage victims being classified as adults. This overestimation of age is also the primary reason why the accuracy in the top age range [18-25] is higher for these services.

| **Range** | **AWS** | **Azure** | **DEX** | **DS13K**<br>(our approach) | **How Old** |
|---|---|---|---|---|---|
| 0-5 | **0.88** | 0.69 | 0.00 | 0.77 | 0.78 |
| 6-10 | 0.43 | **0.66** | 0.13 | 0.44 | 0.49 |
| 11-15 | **0.40** | 0.15 | 0.25 | 0.16 | 0.24 |
| 16-17 | 0.15 | 0.00 | 0.17 | **0.68** | 0.03 |
| 18-25 | 0.87 | **0.97** | 0.89 | 0.70 | 0.95 |
| AVG | 0.550 | 0.496 | 0.293 | **0.553** | 0.503 |

Table 6.5: Accuracy per Group per Service.

Due to the results encountered by the proposed model, and promising figures for an age range which is of interest, because of its proximity to the borderline of adulthood [16-17], it was decided to include the model in the ensemble approach experiment discussed in the next section.

### 6.2.4 Comparison with Ensemble Learning Techniques.

The third experiment was intended to investigate whether ML ensemble techniques can be used to combine all the individual method results as a data fusion approach to improve on the performance exhibited by the existing systems beyond that of each individually.

Given that the existing systems all rely on ML technology, any combination of their results constitutes an ensemble approach. Because the aim of the activity is to compute a predicted age for each subject, regression techniques were considered for this task.

Three standard regression techniques were chosen, namely a logistic regression, gradient boosting and a bagging regressor. These were chosen after observing the results of a number of other regression techniques on this problem. To calculate predicted ages for all of the subjects in the dataset, 10-fold cross validation was used. Here, 90% of the dataset is used for training, with the regressors tasked with predicting ages for the remaining 10%. The training data consisted of the predicted ages for each subject image provided by five systems: AWS, How-Old.net, Azure, DEX and DS13K. This process is repeated 10 times so that the predictions are computed for the entire dataset.

To evaluate this experiment, the results of the regression output were compared to each of the five input systems. This comparison was conducted in two ways: firstly the overall MAE was calculated for each technique, and following this the classification accuracy was calculated for the same age ranges used in the previous section. The MAE for each technique across the entire age range [0-25] is shown in Table 6.6.

| Method | MAE |
|---|---|
| **GradientBoostingRegressor** | **2.425** |
| **BaggingRegressor** | **2.623** |
| **LogisticRegression** | **3.120** |
| AWS | 3.349 |
| DS13K | 3.964 |
| How-Old.net | 5.281 |
| Azure | 5.347 |
| DEX | 6.936 |

Table 6.6: Mean Absolute Error Rates for the 0-25 Age Range.

This table indicates that the three regression algorithms employed achieve a lower MAE than the individual systems. This is an interesting result in that it demonstrates that the regression models that were used reduce the age estimation error when compared with the individual systems. This strongly motivates further research into regression techniques as a promising method to reducing error rates for the facial age estimation problem. Given that the various systems have different performance characteristics across the age range (as evidenced by the results from Section 6.2.2 in particular), these regression models can learn the characteristics of each in order to reduce this effect when combining their outputs.

Given that regression techniques do have a lower error rate than the other approaches within this age range, it is subsequently of interest to find whether their use is also motivated by their performance on the age-range classification task. When the images are divided into age ranges, the accuracy of the regression techniques was also calculated. This did not require a separate experiment to be run; rather an alternative evaluation was conducted. For this evaluation, the important consideration was whether the specific age predicted by the regressor was within the correct age range for each subject. The accuracy of each regressor for each age range is presented in Table 6.7, and compared with the underlying input systems in Figure 6.7.

| Range | Logistic Regression | Gradient Boosting | Bagging Regressor |
|---|---|---|---|
| 0-5 | **0.734** | 0.703 | 0.707 |
| 6-10 | 0.575 | **0.665** | 0.553 |
| 11-15 | 0.432 | 0.391 | **0.441** |
| 16-17 | 0.006 | **0.609** | 0.428 |
| 18-25 | **0.867** | 0.684 | 0.713 |
| AVG | 0.523 | **0.611** | 0.569 |

Table 6.7: Several regression approaches per age range. In each age range, the best accuracy is observed in bold and in average, the best performer is gradient boosting.

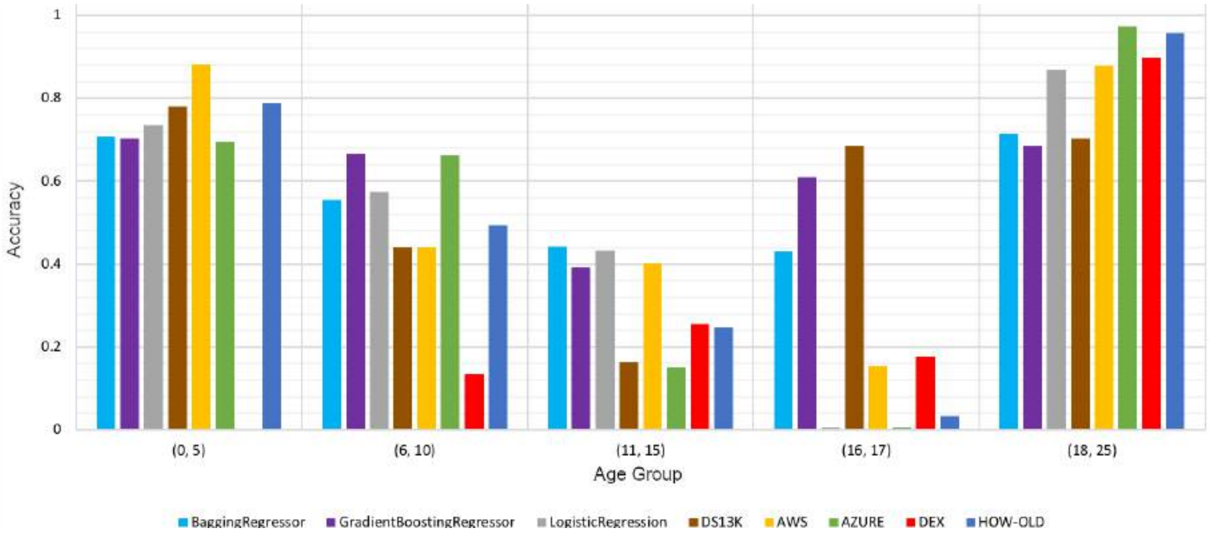


Figure 6.7: Age Estimation Accuracy per Age Group.

From these, it can be seen that the logistic regression, while achieving an overall MAE better than the underlying systems, does not exhibit a promising pattern in terms of the age ranges. Its accuracy in the key 16-17 age range is below almost all other approaches. In contrast, the Gradient Boosting and Bagging approaches both show positive results in this range, with both achieving higher accuracy than the four third-party services that were used.

For underage subjects, the accuracy rates of AWS, How-Old.net and Azure decrease through age ranges as opposed to the adult range [18-25]. It can be observed in Figure 6.7 that most online services have trouble classifying images in the core [16-17] bracket but that both the Gradient Boosting and Bagging ensemble approaches and the DS13K model have much better accuracy in this range.

Given the results in the previous sections, it is unsurprising that AWS, How-Old.net and Azure have the poorest performance for underage subjects near the borderline. In Section 6.2.2, they are shown to generally overestimate a subject's age in this range, thus frequently misclassifying them as adults. Furthermore, the results in Section 6.2.2, specifically Figure 6.6 indicate that their MAE/Year is greater from the region 13 to 19 years of age in the dataset. Unsurprisingly, the classification accuracy reduces as underage ages get closer to the cut-off point of 18. For 17 year old subjects, DEXs MAE/Year is the lowest, meaning that the performance is better for that particular age than the rest of the services, whereas Azure has the worst performance between them. Their tendency to overestimate ages results in higher accuracy figures for overage subjects. An 18 year old is very rarely (less than 10% of the time) misclassified as being underage.

On the other hand, the accuracy of the regression models is much higher than for the underlying systems when averaged over the age ranges. Overall, the Gradient Boosting approach shows the best results. Even for 17 year old subjects, it has a better performance over the rest of ensembles, though failing to beat the DS13K model.

One notable finding is that the ensemble approaches have lower accuracy for subjects who are equal and over 18. This is partially due to the tendency of the underlying systems to overestimate ages, which will naturally lead to high accuracy for overage subjects in the highest age bracket. However, the accuracy of the regression models for overage subjects is far in excess of the accuracy figures for the underlying systems for underage subjects. This is closely related to their overall lower error rates within this age range.

When evaluating this result, it is also important to keep in mind the use cases for these technologies. Arguably the consequences of misclassifying a younger subject as being overage are much more serious than the opposite scenario. If these systems are to be used in a forensic scenario to automatically identify potential victims of child abuse, it is important that such victims are not missed by these systems. Wrongly classifying a child as being older may result in a case not coming to the attention of investigators. In contrast, erroneously allocating an older subject as being younger may ultimately result in wasted investigator effort to examine a situation that is ultimately non-criminal. There is a strong argument to be made that the latter event is much less serious. Even in this scenario, a false positive classification of an adult subject as being underage would trigger a manual evaluation, thus placing investigators in the same position as if the technology was not used.

However, given the multi-year backlog in conducting digital forensic investigations in many jurisdictions [156], clearly an approach that improves accuracy overall is desirable. While the results presented in this section show great promise, it is clear that further work is required to improve the performance of facial age identification even further if it is to be adopted on a wide scale as part of digital forensic investigators' toolkits.

## 6.3 VisAGe Dataset Generation Process

In this section we discuss the voting statistics that took place in the dataset generation process. Although the process is assisted by automation, humans take part in the input data and sometimes make mistakes. Further, once the dataset was created, the characteristics and data were extracted such as sample data, distribution, physiological predicted variables, quality of the image, makeup/facial accessories, facial composition, etc. Finally, the amount of human voting efforts is analysed and presented.

### 6.3.1 Voting Statistics

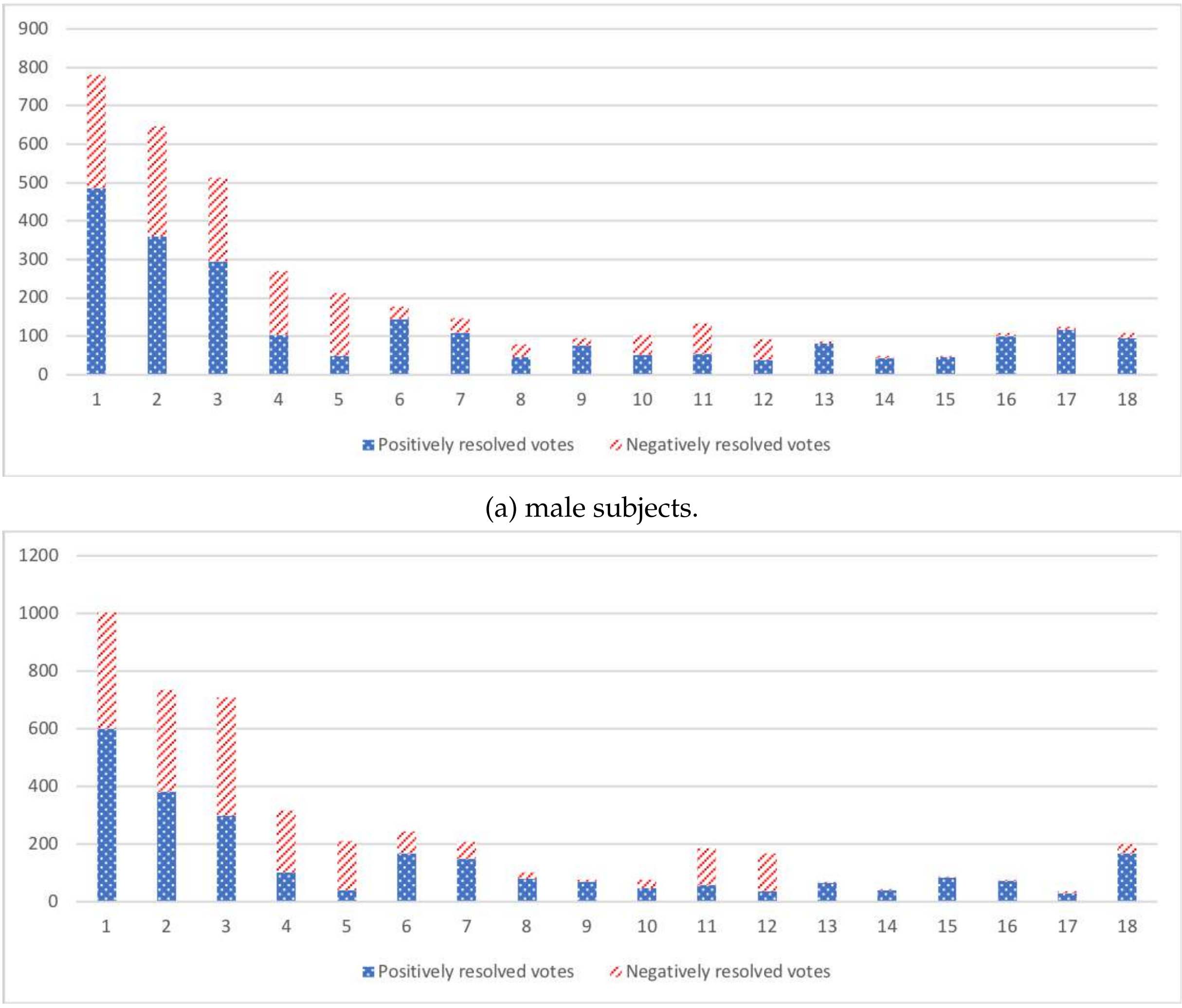
A binary vote only allows a single answer from two possibilities. For instance, Positive/Negative. For a vote to automatically be positive, three positive votes are required. Similarly, for a vote to be automatically discarded, 2 negative votes are required. When the votes have been completed, if there is a disagreement but there are 2 positive votes, the image is analysed for a second time by an Inspector.

As per Figure 4.9, there are three actors: Voter, Inspector and Q/A Agent. The voting procedure is prone to human error or subject to disagreement; therefore, the Voter has either made a mistake or the vote was qualified as disagreed. The Inspector is the actor that reviews the disagreed images and then resolves the votes to either positive or negative. In Figure 6.8a it can be seen that the majority of male disagreed votes happen in the early ages (1, 2 and 3-year-olds). It can also be seen depicted that negatively resolved votes are predominant at ages 4 and 5. The same behaviour can be seen in graph 6.8b which corresponds to female disagreed votes.

All the actors are able to delete images that are considered inappropriate for the dataset. Nevertheless, the main actor that deleted most images was the Q/A Agent. In Figure 6.9 it is shown that the most deleted photos are for 1 year-old subjects. This being due to the age group having the most occurrences of images through the entire age range as depicted in Figure 6.10.

### 6.3.2 VisAGe Dataset Characteristics

The distribution of the VisAGe dataset by age and gender is outlined in Section 6.3.2.1. A sample of the dataset is also presented by way of an average image per age per



(a) male subjects.

(b) female subjects.

Figure 6.8: Number of Positively and Negatively Resolved Votes per Age Bracket

gender. The “shape” of VisAGe distribution is down to what was available using the image acquisition methodology. Quality of the images are explored in Section 6.3.2.3 where the resolution, occlusion, blur, noise and exposure of the images are analysed. Section 6.3.2.4 appraises several artificial attributes that have been detected using Microsoft Azure Cognitive services such as the presence of makeup, glasses and other accessories. Furthermore, the natural facial compositions of the subject are examined in Section 6.3.2.5, where the subject’s facial expression and facial hair are taken into account.

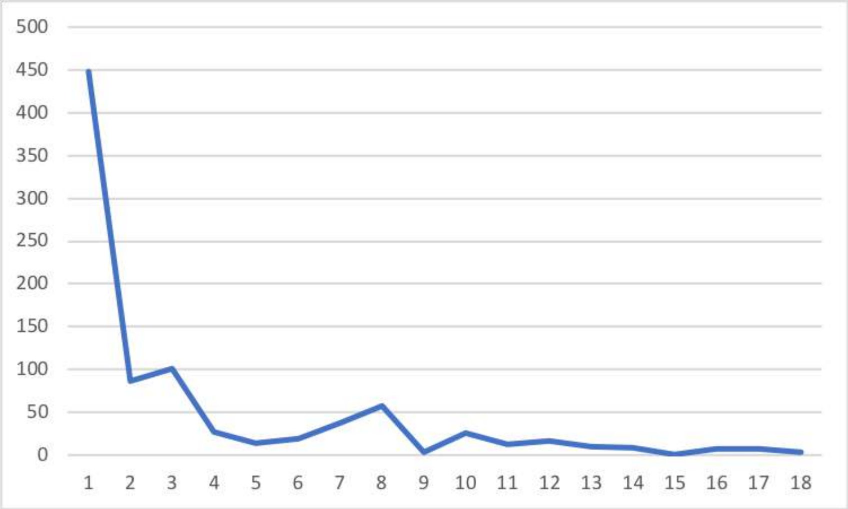


Figure 6.9: Number of Deleted Photos per Age of Subject.

#### 6.3.2.1 Dataset Sample and Distribution

An overview of the VisAGe dataset distribution is illustrated in Figure 6.10. The VisAGe dataset is created through the use of Flickr's repository of CC photographs. Hence, Flickr's availability for the different age ranges has contributed greatly towards the distribution of the dataset. The strength of the dataset lies predominately in the early stages of youth, where a much denser distribution of records can be observed. The majority of the dataset is comprised of the 1- to 4-year-old age range with image numbers exceeding 1,000 for both female and male categories. A much thinner dispersion of data, however, is present at the teen age range with 12 and 15-year-old males and 14 and 17-year-old female having the least amount of images available (less than 100 records).

A sample of the dataset can be located in Figure 6.11 as an average image per class per gender from the age range 1 to 18. It can be deduced that whilst the colouring of the subjects clothing is an unreliable means of analysis to distinguish the gender of the image, we can still identify that the colour blue is prominently worn by male minors particularly for 1, 3, 9, 10, 17 and 18-year-old's. This observation is expected if emphasis is given on the stereotypical view that blue is for male children[203].

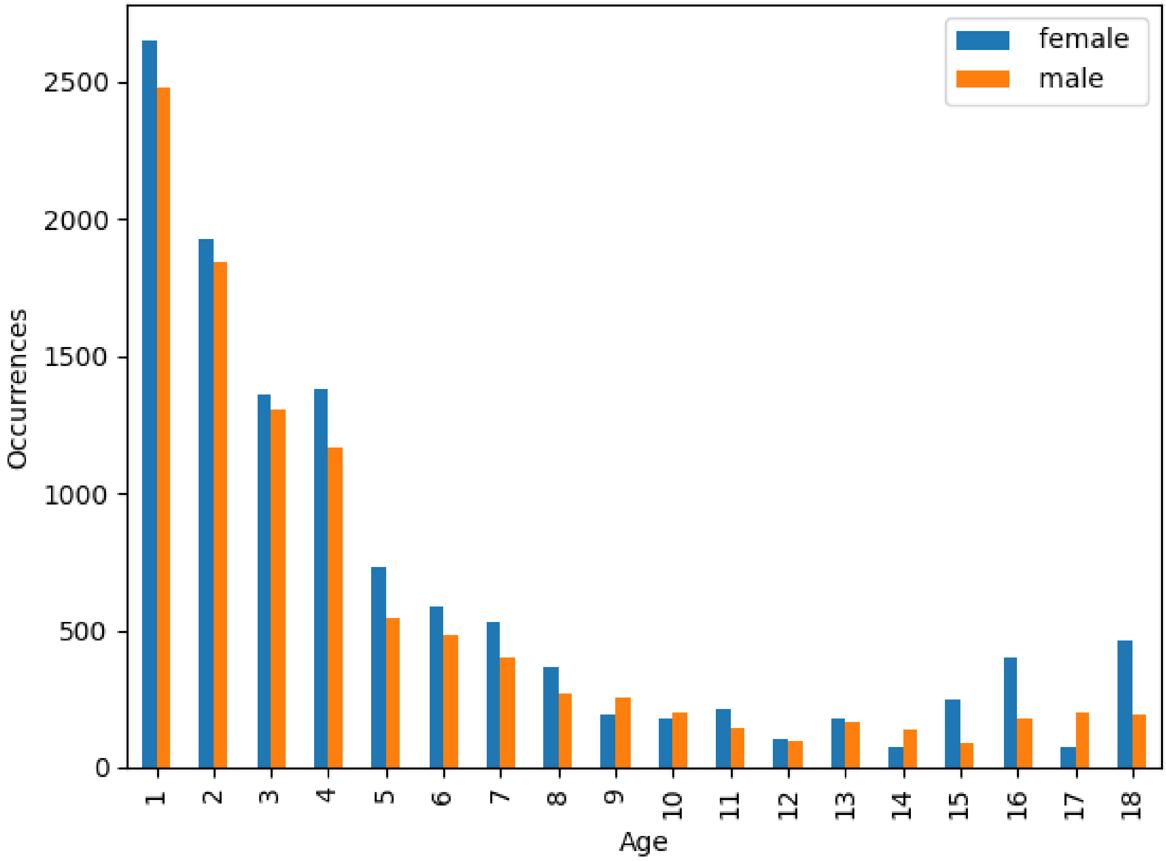


Figure 6.10: VisAGe Dataset Age Distribution by Gender.

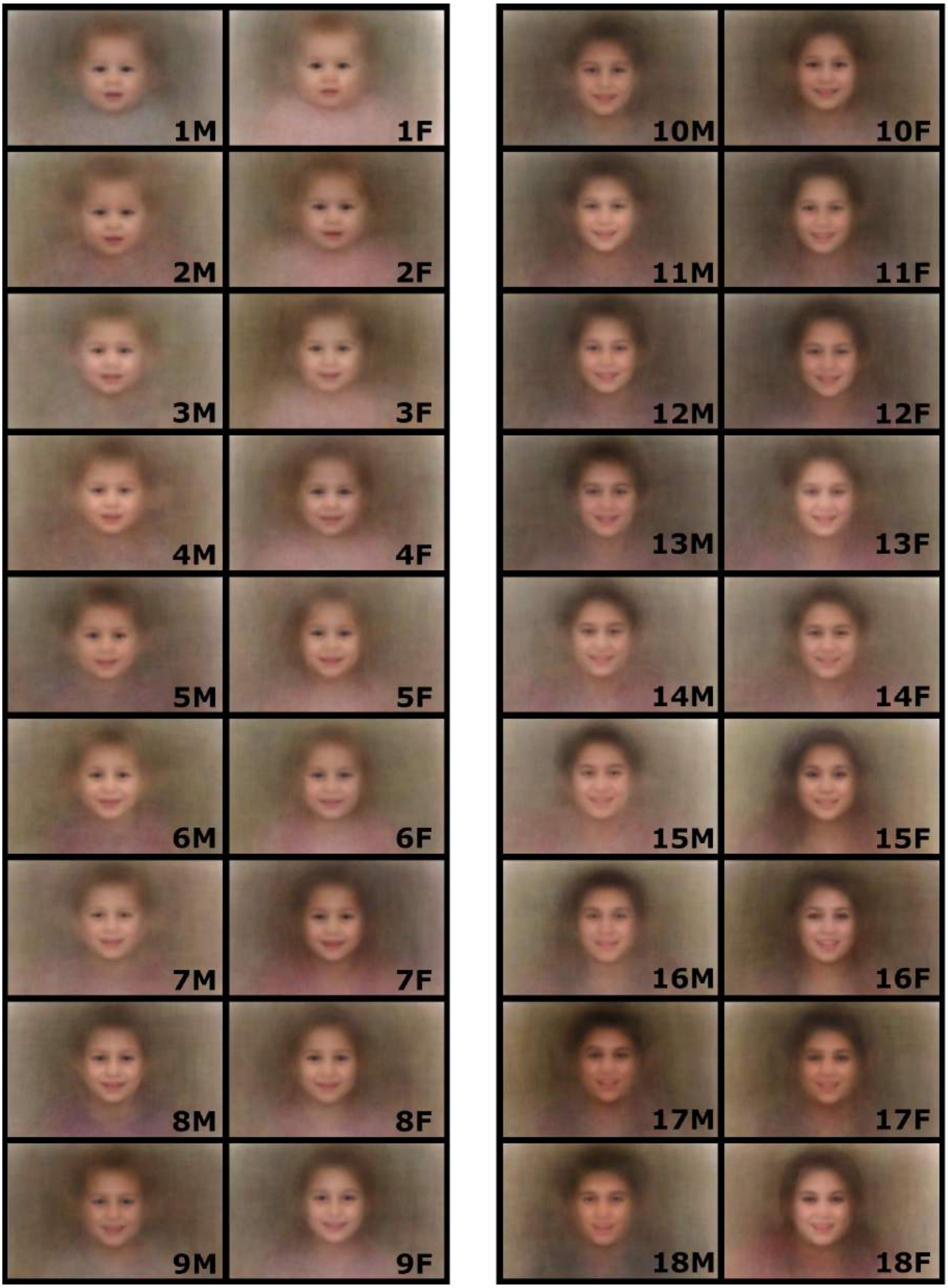


Figure 6.11: Image Average VisAGe.

#### 6.3.2.2 Physiological Predicted Variables

Physiological variables such as age and gender were predicted with the Microsoft Azure Face API. The age field previously mentioned in Table 4.5 represents an estimated “visual age” number in years. It depicts how old a person looks like rather than the actual biological age. The Azure age performance for underage subjects was evaluated in Section 6.2.2 with a recorded MAE of 5.347 for underage subjects tested over a sample of approximately 1,270 images.

The predicted gender was used to separate groups of male and female subjects. Nevertheless, an experiment to measure gender inaccuracies between annotated gender and Azure predicted gender in the VisAGe dataset was accomplished and a percentage of 23.29% failure was obtained. The rate was calculated by the number of wrongly classified gender over the size of the dataset.

#### 6.3.2.3 Quality of Images

Image resolution is defined as the fineness of detail that can be clearly distinguished in an image [19]. It is composed by height and width and the overall image dimensions statistics can be seen depicted in Table 6.8.

A plot of the mean height and width of an image per age is described in Figure 6.12. It is observed that there are images with width and heights over 7,300 pixels. The average images are 2,317.75 x 2,571.87 pixels, this measurement equates approximately to 6 Megapixels. The quality of the dataset can also be measured by several factors such as the ones depicted in Table 6.9: occlusion, blur, noise and exposure.

| Statistic | Height | Width |
|---|---|---|
| mean | 2,317.75 | 2,571.87 |
| std | 1,140.94 | 1,348.18 |
| min | 112.00 | 150.00 |
| 25% | 1,426.00 | 1,500.00 |
| 50% | 2,304.00 | 2,482.00 |
| 75% | 3,024.00 | 3,456.00 |
| max | 7,712.00 | 7,360.00 |

Table 6.8: Statistics of Image Dimensions (Height and Width) in Pixels.

Images within the dataset were validated and quality assurance of the images were performed, thus, the occlusion characteristic of the images were predicted to be close to 0. Three main types of occlusion were investigated; mouth, eye and forehead. As forecast, the overall percentage of occlusion present in the images were found to be

| Attribute | Type I | % | Type II | % | Type III | % |
|---|---|---|---|---|---|---|
| occlusion | mouth | 2.22 | eye | 0.50 | forehead | 3.69 |
| exposure | good | 87.35 | over | 10.77 | under | 1.86 |
| blur | low | 73.91 | medium | 21.70 | high | 4.37 |
| noise | low | 70.09 | medium | 22.23 | high | 7.66 |

Table 6.9: Microsoft Azure Face Application Programming Interface Quality Attributes.

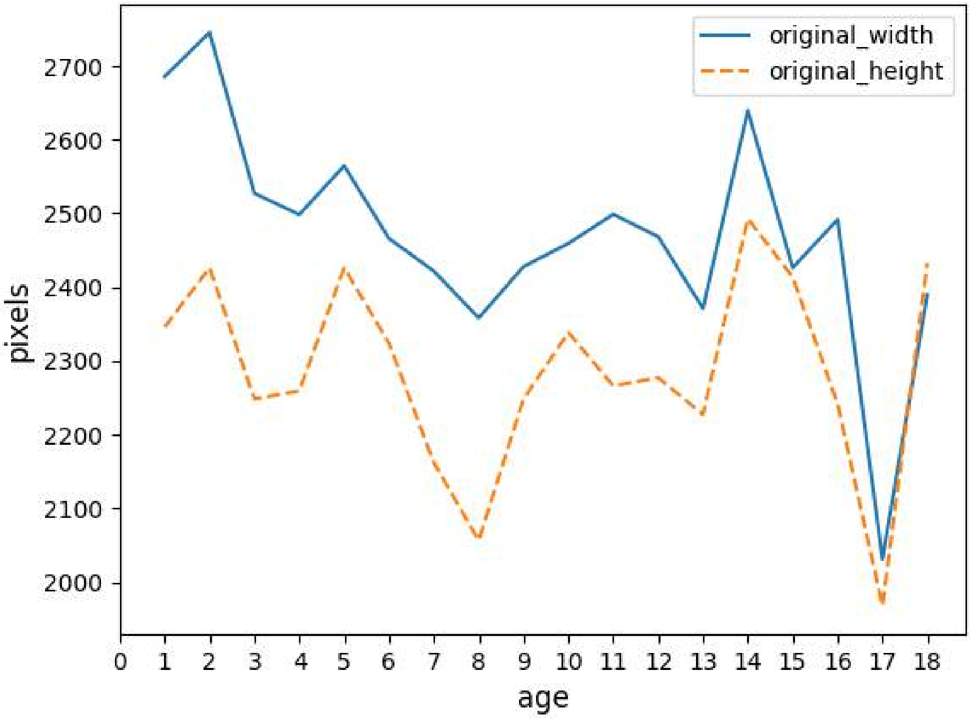


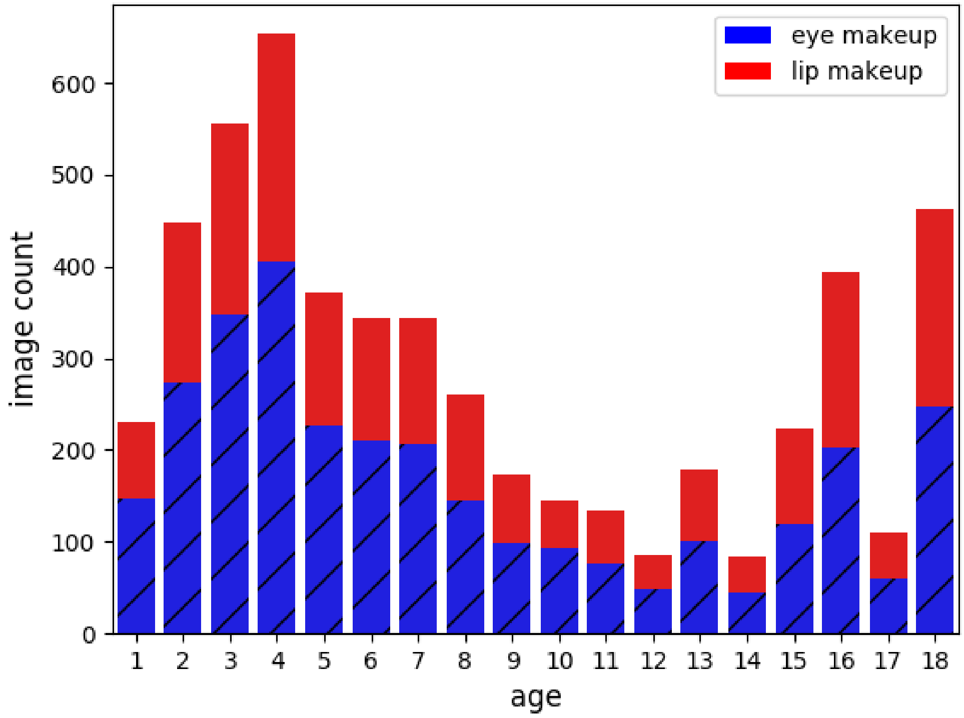
Figure 6.12: Mean Height and Width per Age.

relatively low with 2.22%, 0.50%, 3.69% for mouth, eye and forehead occlusion respectively. Similarly, a total of 87.35% of the images were found to be of good exposure with only 10.77% and 1.86% of over and under exposure respectively. Another favourable result was also obtained for the blur factor; Only 4.37% of the images were detected to have a high-blur factor with the majority, 73.91% having low-blur factor. Lastly, the noise present in the images were broadly low, 70.09% demonstrated low quantities of noise with only 7.66% of the images experiencing high level of noise. For the full summary of the results obtained for the quality of the images, please refer to Table 6.9. The aforementioned measurements were obtained through the use of Microsoft Azure Face API.

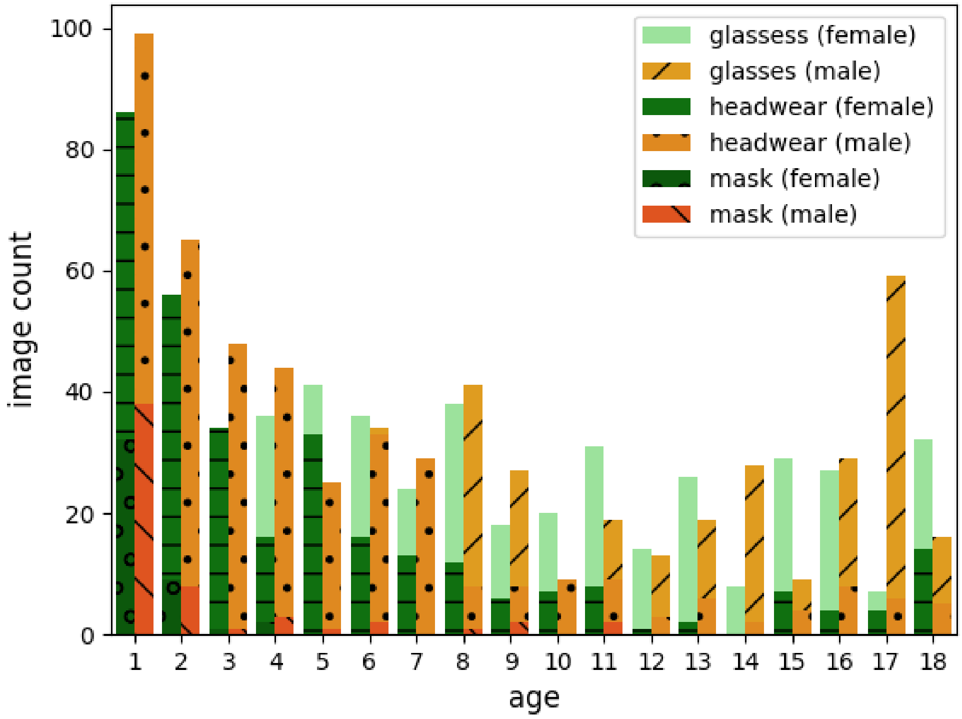
#### 6.3.2.4 Makeup and Accessories

There are occasions when subjects make use of makeup and accessories. Despite the focus towards underage subjects, considerable percentages of facial makeup were found. It was detected that the percentage of makeup usage for both lips and eye amounted to 37.74. As depicted in Figure 6.13a, the ages with most exposure to makeup were 4, 16 and 18-year-olds. In the early ages it can be seen that there is a significant amount of images that are perceived to have makeup. This may be due to the detection being triggered by food colouring that has been transferred to the subject's facial area, especially in birthday parties.

Facial accessories occurred relatively infrequently. The sum of counts for glasses, headwear and masks were equivalent to 7.41%. The glasses accessory had a 3.93% presence in the dataset. The details can be seen in Figure 6.13b. From 4 years on-wards, glasses were present in each age range for female subjects. However, for males, the glasses appeared from 8 years on-wards but was not present in each age class. Furthermore, it can be observed that the headwear accessory had the most occurrence in a given age, 1-year-old subjects. Both male and female counterparts for this age had the most subjects displaying head-wear. Conversely, the mask accessories were found to be the least frequent accessory classification type. In general, the majority of its occurrence were found primarily in ages 1 and 2. Overall, we can deduce that accessories were detected predominantly in 1-year-old subjects. The amount of detected accessories decreases gradually per age until the 12 years-old mark point at which the amount of accessories start to rise again.



(a) Stacked Makeup (Eye and Lips)



(b) Count of Accessories (Glasses, Headwear and Mask)

Figure 6.13: Makeup and Accessories per Age Categories

#### 6.3.2.5 Facial Composition of the Subject

Amongst artificial objects that can conceal and distort the subject as discussed in 6.3.2.4, natural occurring components can also affect the subject such as the presence of facial hair, facial expression and positioning of the head.

Whilst the dataset contains images of minors, several images have been detected to possess subjects with varying amount of moustache, beard and sideburns for both genders. It can be ascertained in Figure 6.14 and through manual observation of the images that were afflicted, that the accuracy of the Microsoft Azure Face API tool was not always accurate. It was found that under certain circumstances, images that incorporated some form of occlusion around the mouth and sides of the face such as a mask, food around the face or a shadow, have been incorrectly detected as facial hair.

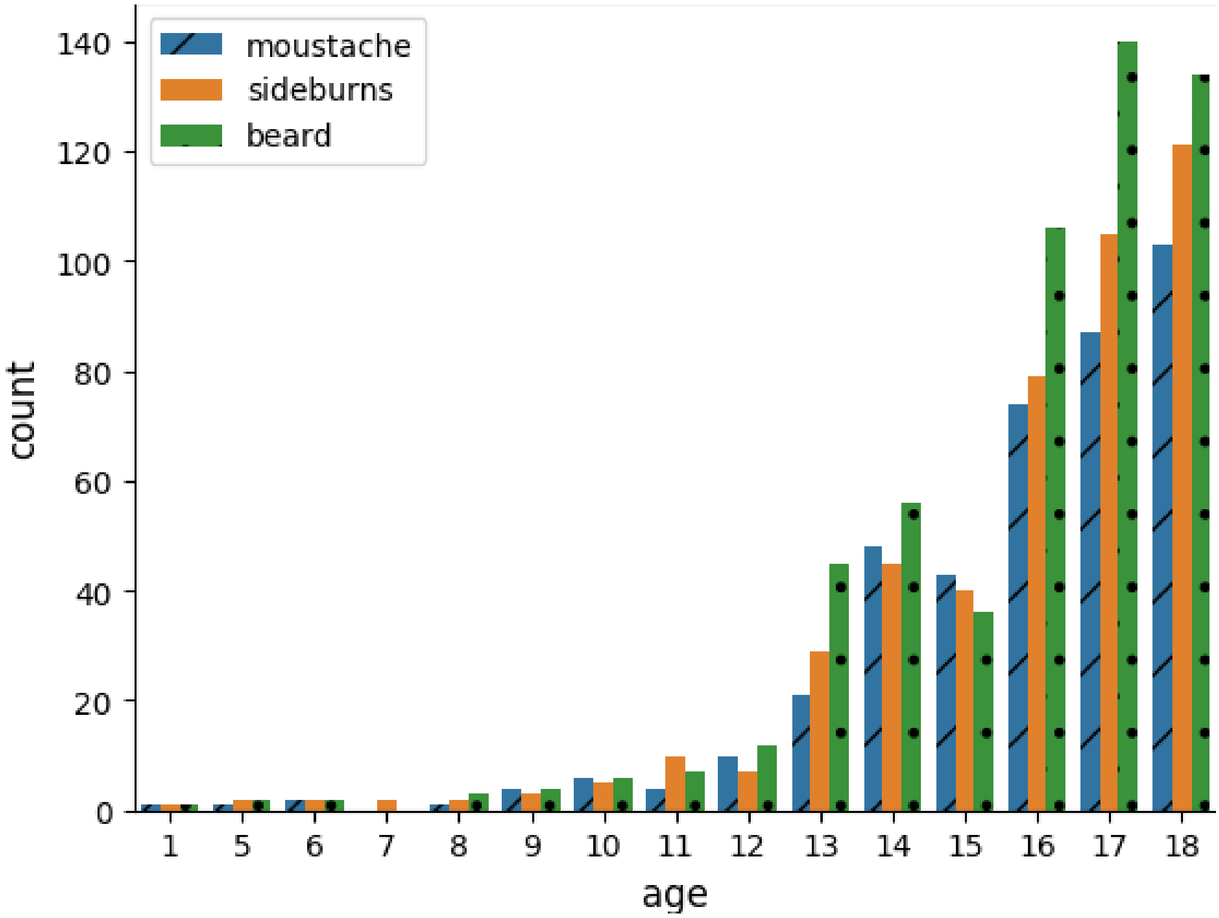


Figure 6.14: Predicted Male Facial Hair Counts per Age.

With the use of Microsoft Azure Face API, a variety of emotions were detected within the VisAGe dataset: neutral, anger, contempt, disgust, fear, happiness, sadness and surprise. The bulk of the subjects were found to have both happy and neutral sentiments as shown on Figure 6.15. On the other hand, an extremely small fraction of the data was found to have the feeling of contempt, disgust, fear and anger. These results were expected as the method employed in gathering the images was based on gathering images from birthday events which can be considered as happy occasions.

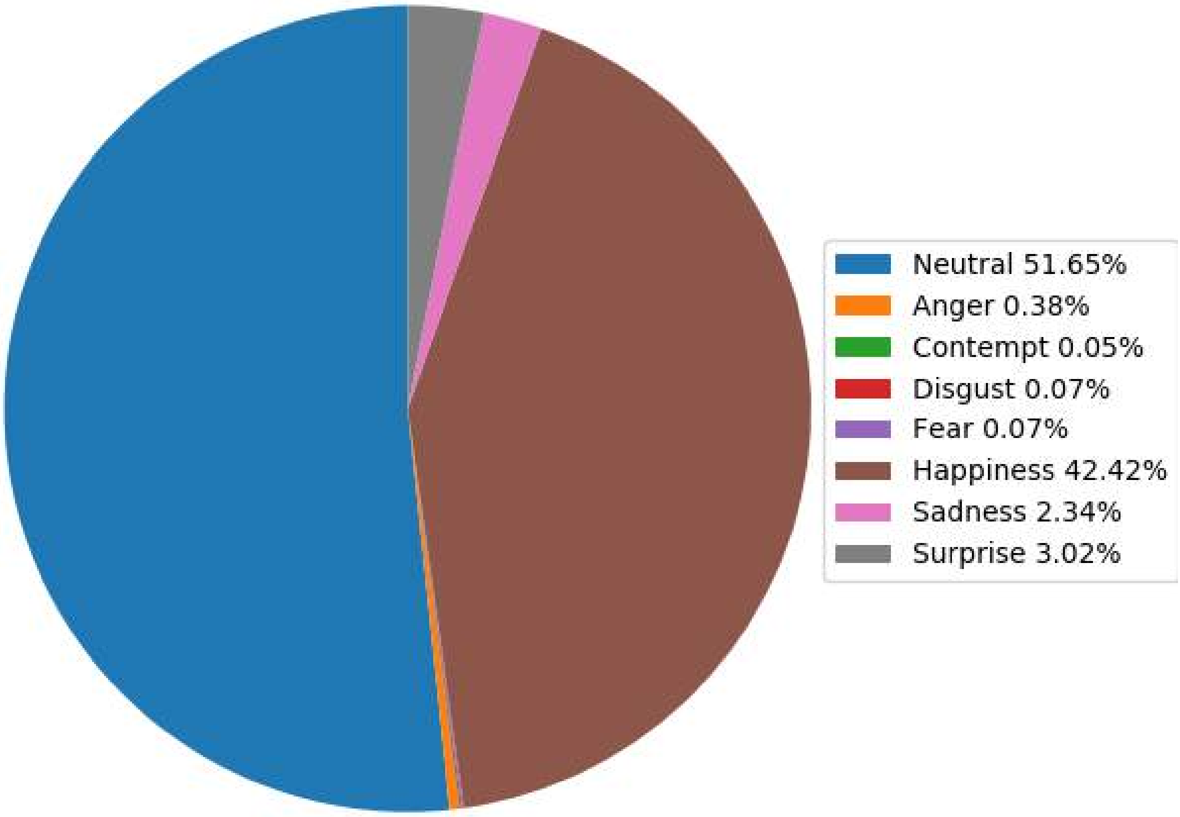


Figure 6.15: General Distribution of Emotions.

In Figure 6.16, a rising smile intensity can be seen observed. In general, female subjects tend to smile with more intensity than male subjects, with an exception of 8 and 15 year-old males. The peak of intensity for males and females is 8 and 14 years-old respectively.

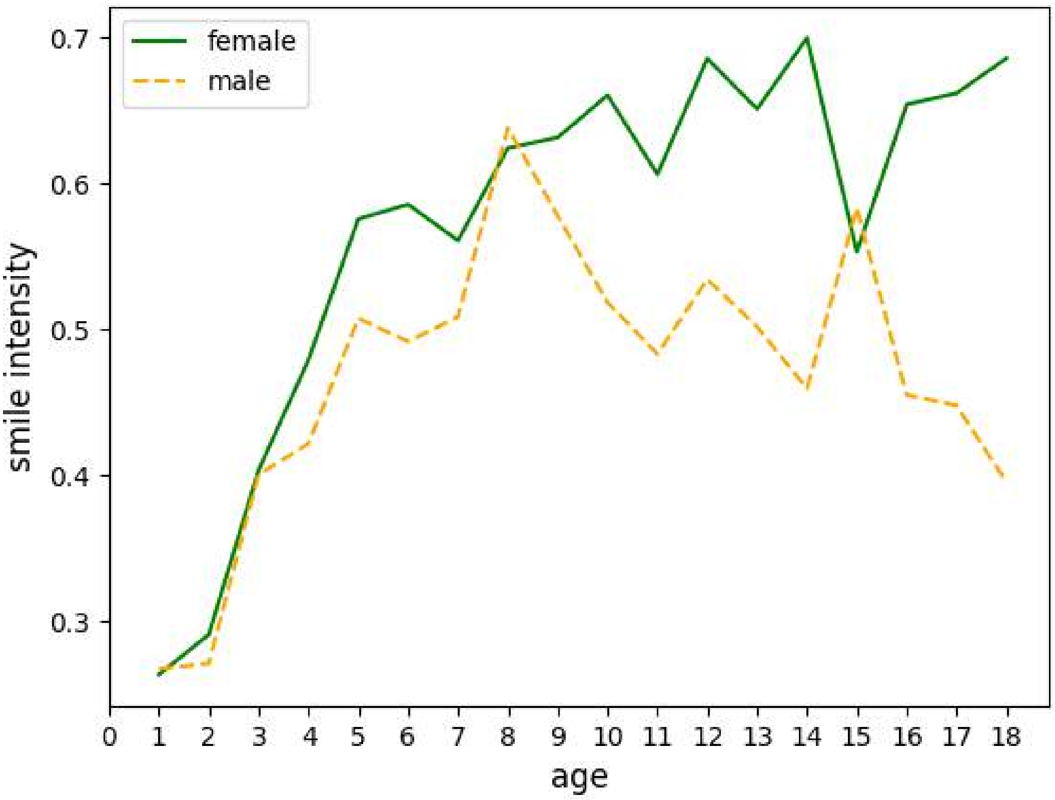


Figure 6.16: Mean Smile Intensity per Age per Gender.

Head pose can be a social signal and can be determined by three degrees of freedom: yaw, roll, and pitch. The measured mean head pose angle is variable for yaw and roll within each age group whereas the mean pitch angle value tends to be close to 0. This can be seen in Figure 6.17 and would have been visible in Figure 6.11 if no face alignment was implemented. For the whole dataset, the statistics are shown in Table 6.10. Mean values of -0.72, 0.07, -0.01 are observed for Yaw, Roll and Pitch angles respectively.

| **Statistic** | **Yaw** | **Roll** | **Pitch** |
|---|---|---|---|
| mean | -0.72 | 0.07 | -0.01 |
| std | 12.07 | 10.24 | 0.82 |
| min | -66.00 | -47.00 | -29.90 |
| 25% | -7.30 | -5.60 | 0.00 |
| 50% | -0.40 | 0.00 | 0.00 |
| 75% | 6.30 | 5.80 | 0.00 |
| max | 63.20 | 50.70 | 20.40 |

Table 6.10: Statistics of Head Pose.

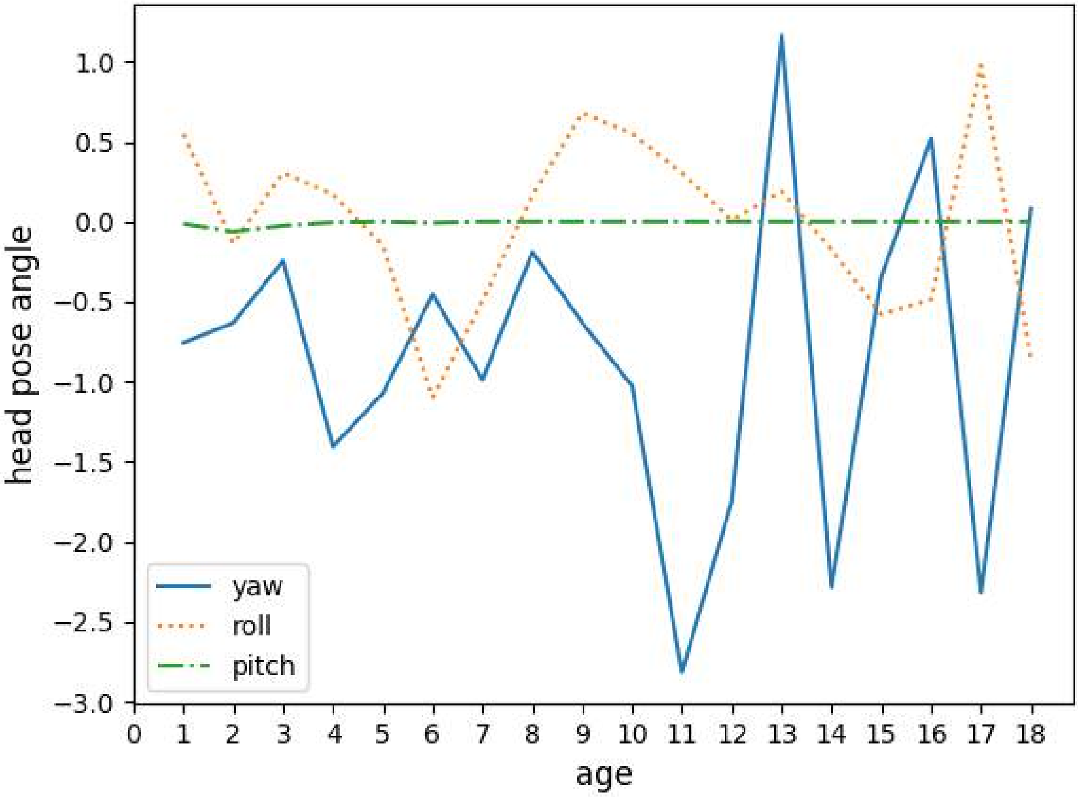


Figure 6.17: Mean Head Pose Angle per Age.

Finally, hair distribution has been analysed. The majority of subjects have hair (94.66%) with a low percentage of non-visible hair and bald subjects with 3.57% and 1.76% respectively. Details of hair colour distribution can be seen in Figure 6.18. It is clear that there is a great number of blond subjects in the early years and brown haired subjects are predominant throughout the entire age range.

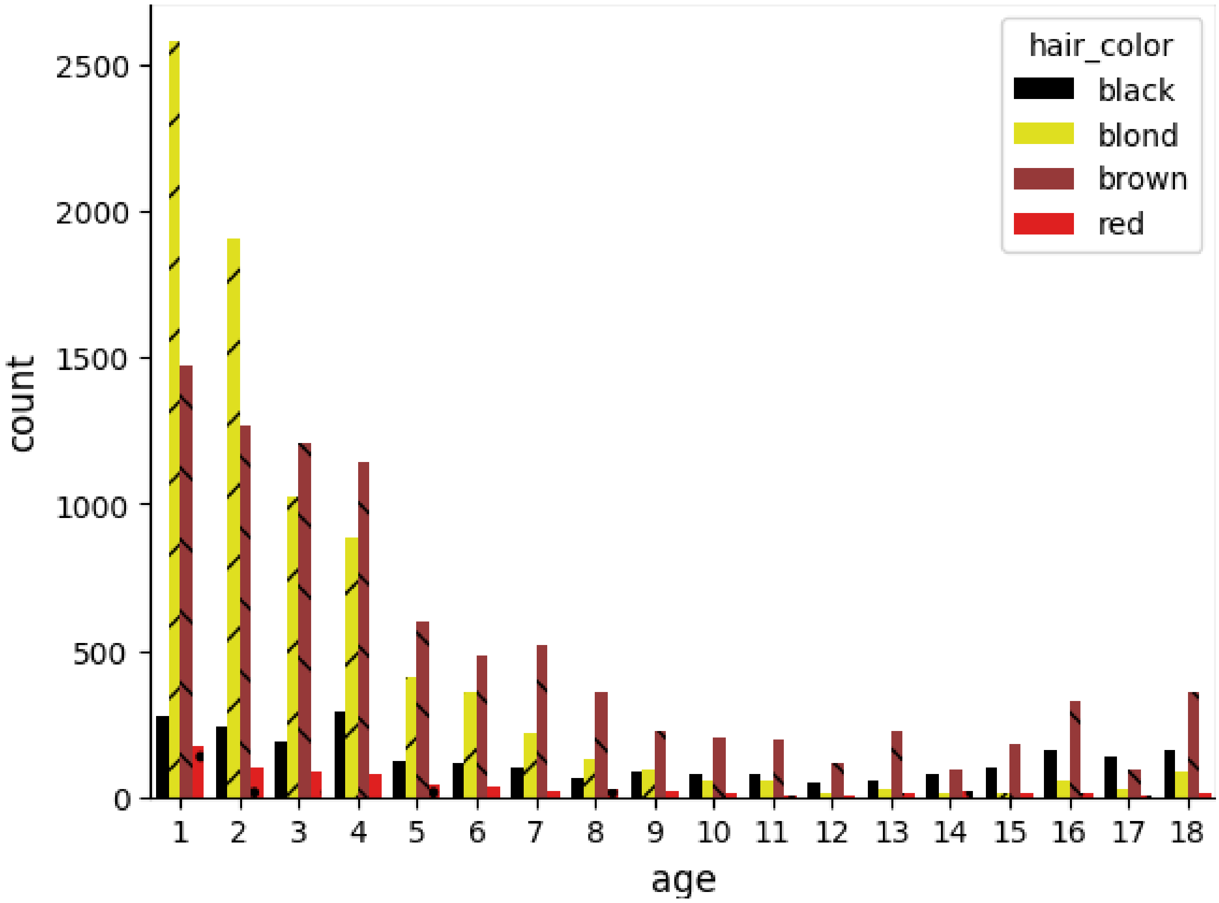


Figure 6.18: Counts of Hair Colour per Age.

### 6.3.3 Voting Human Effort

Each voting was recorded in an *ObjectId* format as previously explained in Section 5.3.1.2. From this format the date can be extracted. The votes were grouped by day and the minimum and maximum time were obtained. The number of hours worked by day by a user were computed by the difference between the minimum and maximum hours. The voting effort per day per user can be seen depicted in Figure 6.19. The first votes can be seen recorded on the 1st of April 2018. The last votes can be seen on the 1st of January 2020. This amounts to a time span of 1 year 9 months approximately. The number of users was 16 and the effort to complete the voting was around 1,170.17 hours.

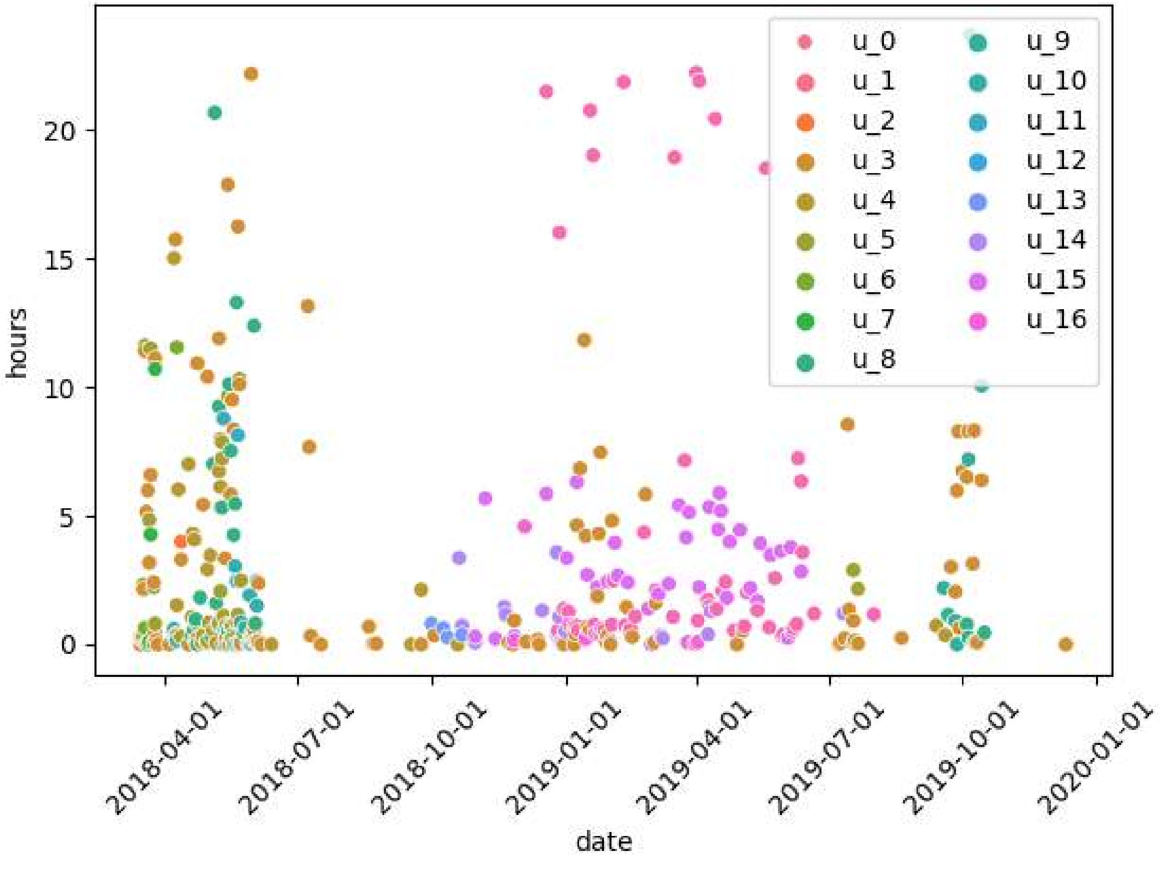


Figure 6.19: Total of Voting Hours per Date per User

## 6.4 Age Estimation Influencing Factors

Machine annotated attributes have been assessed from the two most prominent cloud-based services. Linear correlations linking the MAD between real age and predicted age have been evaluated to detect the age estimation influencing factors. These have been categorised into weak, mild and strong accordingly to the correlation. Due to the different rates of performance, the two cloud services have been assessed independently. Overall, Microsoft Azure achieves a MAE of 2.082 for the VisAGe dataset, whilst AWS has a MAE of 4.075. Furthermore, the distribution of $\mathrm{Er_d}$ for each class service has been analysed. It must be noted that for all succeeding correlation figures, the attribute error is shown to have a positive perfect degree of correlation to $\mathrm{Er_d}$ (Refer to Equation 4.4). This is expected as any attribute examined with itself produces this behaviour.

### 6.4.1 Microsoft Azure

Influencing factors affecting Azure's facial age estimation have been evaluated. Section 6.4.1.1 looks into the distribution of correlations between the $\mathrm{Er_d}$ and the pre-defined machine-extracted attributes in order to identify the influencing factors and their gravity towards the $\mathrm{Er_d}$. The distribution of significant correlations of greater than or equal to 5 between attributes are outlined in Table 4.5 and the $\mathrm{Er_d}$ for different ages are represented in Figure 6.21.

#### 6.4.1.1 Strong PCC Distribution per Age with Error Difference Greater Than or Equal to 0

The distribution of strong correlation values have been evaluated per age between $\mathrm{Er_d} \geqslant 0$ and the machine-detected attributes. It was observed that one year old's were the only age that demonstrated any linear correlations. These positive strong correlations were produced by the facial hair attributes: moustache, beard and sideburns. It was anticipated that the presence of facial hair will hinder accurate estimation of facial age. However the cause of facial hair being detected for one-year-old's was produced by incorrect detection of moustaches and beards (typically from food around the subject's mouth). Furthermore, no attribute was identified to be of strong influencing factor towards the accuracy of the age estimator for all succeeding ages, when the $\mathrm{Er_d} \geqslant 0$ is considered.

#### 6.4.1.2 Error Distribution

Figure 6.20 is the uni-variate distribution of observations of the $\mathrm{Er_d}$ value. It can be concluded that the general consensus of Azure's underage facial age estimation is reasonably accurate, i.e., the majority of scores obtained were relatively low with the bulk of the result being less than or equal to 5. It can be observed that there is a great difference on the amount of results achieving accuracy of $\mathrm{Er_d} < 5$ versus larger $\mathrm{Er_d}$ values of greater than or equal to 5. The distribution of strong correlations achieved in Section 6.4.1.1 was further filtered by $\mathrm{Er_d} < 5$ and $\mathrm{Er_d} \geqslant 5$, as discussed in Sections 6.4.1.3 and 6.4.1.4 respectively.

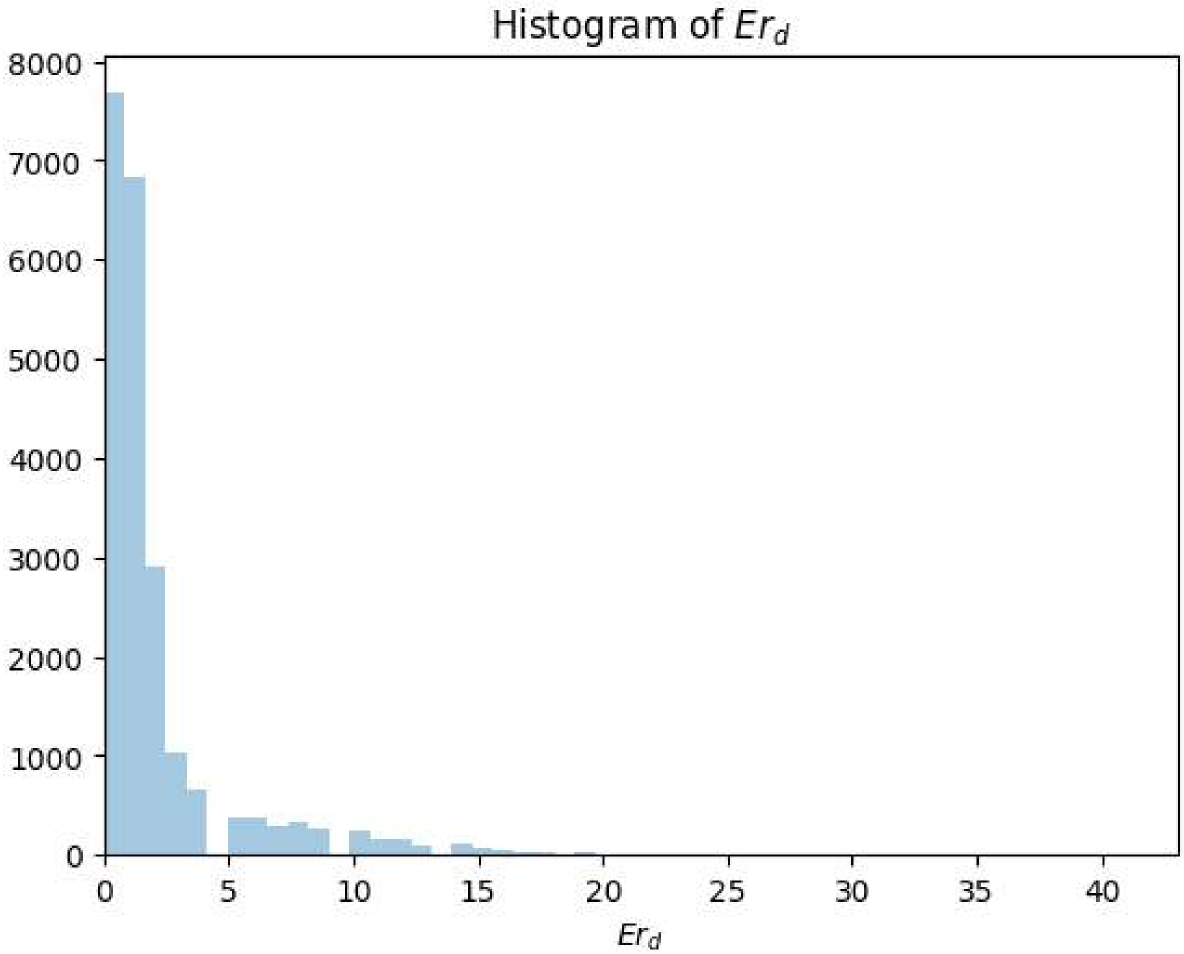


Figure 6.20: Distribution of $\mathrm{Er_d}$ per age using Microsoft Azure Face API predictions.

#### 6.4.1.3 Strong PCC Distribution per Age With $\mathrm{Er_d}$ Less Than 5

Whenever Azure's facial age estimation demonstrates a high level of accuracy, achieving error margins $\leqslant 5$, the distribution of $|PCC| \geqslant 0.5$ presented no correlating data attributes across all ages. These results were similar to that obtained in Section 6.4.1.1. It can be concluded that no influencing factors have been identified to be associated with the estimator achieving good results.

#### 6.4.1.4 Strong PCC Distribution per Age With $\mathrm{Er_d}$ Greater Than 5

Conversely, when the accuracy of the estimator declines beyond the error margins of 5, the distribution of strong correlations have been identified between attributes and the estimator's $\mathrm{Er_d}$ occurring on ages 1, 2, 4 to 7 and again on 9 to 10 years old, as shown on Figure 6.21. No distribution of strong correlation was detected for ages 3, 8 and 11 to 18 where $\mathrm{Er_d}$ is set to greater than 5. To delve further into identifying the attributes triggering these results and by what magnitude, Figures 6.22, 6.23, 6.24 and 6.25 outline the distribution of the aforementioned PCC values according to specific attributes for each age.

For one year-old subjects, an interquartile range (IQR) of strong correlation values around 0.5 to 0.75 were detected, as shown in the Figure 6.21, with outliers lying in the negative region. Figure 6.22 confirms that this outcome was the result of the age displaying strong correlations of 0.81 for *underExposure*, *noiseLevel_medium*, *sideburns, moustache and beard*. These noted attributes, as shown in Table 6.11, were found to have a strong linear influence to the decline in the accuracy of the Azure's age estimator for one year old's. Equally, attributes that displayed strong negative correlations of -0.55 and -0.81 for *noiseLevel_low and goodExposure* respectively, were found to have linear influence in the improvement of estimator's performance accuracy. Furthermore, these two negative attributes contributed to the outliers in the data for age one. Such strong PCC values obtained in the age were not expected as a more diverse set of PCC figures were thought to be more probable.

In Figure 6.21, a similar IQR has been found for 2 year old's. This IQR lies just above the 0.25 to 0.75 range denoting that attributes with PCC values $\geqslant 0.5$ were close to the 0.5 benchmark. By referring to Figure 6.22, it is confirmed that strong correlating attributes had a magnitude of 0.52 and 0.51. The former value corresponds to the manual annotated female gender attribute *_gender_female*. The latter corresponds to the contempt expression. It is observed that for 2 year-old subjects, the manual gender annotation for females had a stronger correlation than it's machine-annotated version, with age

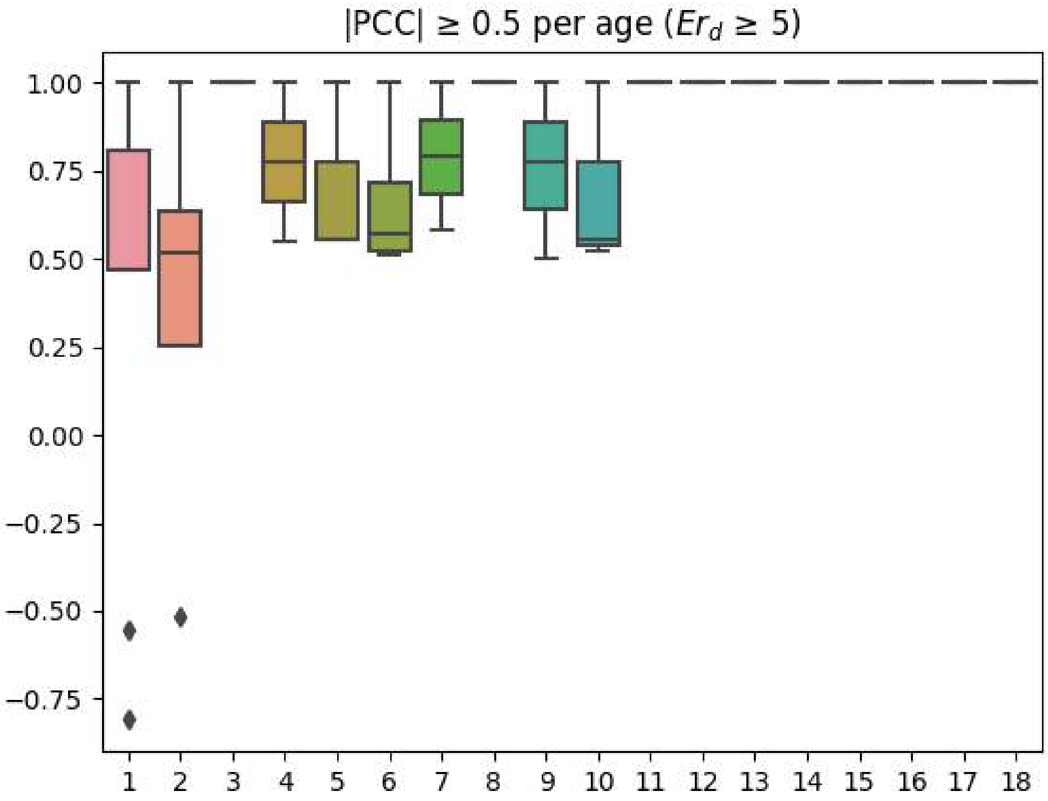


Figure 6.21: Azure: Box-plot of PCC Distribution per Age, where $\text{Er}_\text{d}$>5.

estimation accuracy. In comparison to the preceding age, two year old's presented with more diverse assortment of attributes, only 3 of the attributes managed to achieve strong correlation values of over 0.5 in magnitude; these strong influencing attributes ($|PCC| > 0.5$) are outlined in Table 6.11. Gender was the key prominent attribute that influenced the increase and decrease of $\text{Er}_\text{d}$ for two year old's; female subjects caused a decline in accuracy for 2 year old's, whilst male subjects were found to linearly influence the incline of the accuracy. Additionally, emotion of contempt was also found to be a strong influencing factor affecting the accuracy for two year old's. All succeeding ages, as shown on both Figures 6.21 and 6.26, present no strong negative correlation above the -0.5 threshold. Therefore no influencing factors have been identified that elevate the gravity of the $\text{Er}_\text{d}$ for ages 3 and above. Moreover, for 4 year old's, Figure 6.22 illustrates only one attribute to have strong association with the accuracy of the age estimator; emotion of anger with value 0.55.

Ages 5, 6, 9 and 10 all exhibit forms of facial hair correlations with the performance of Azure's facial age estimation on the underage dataset (again, miscategorisations at

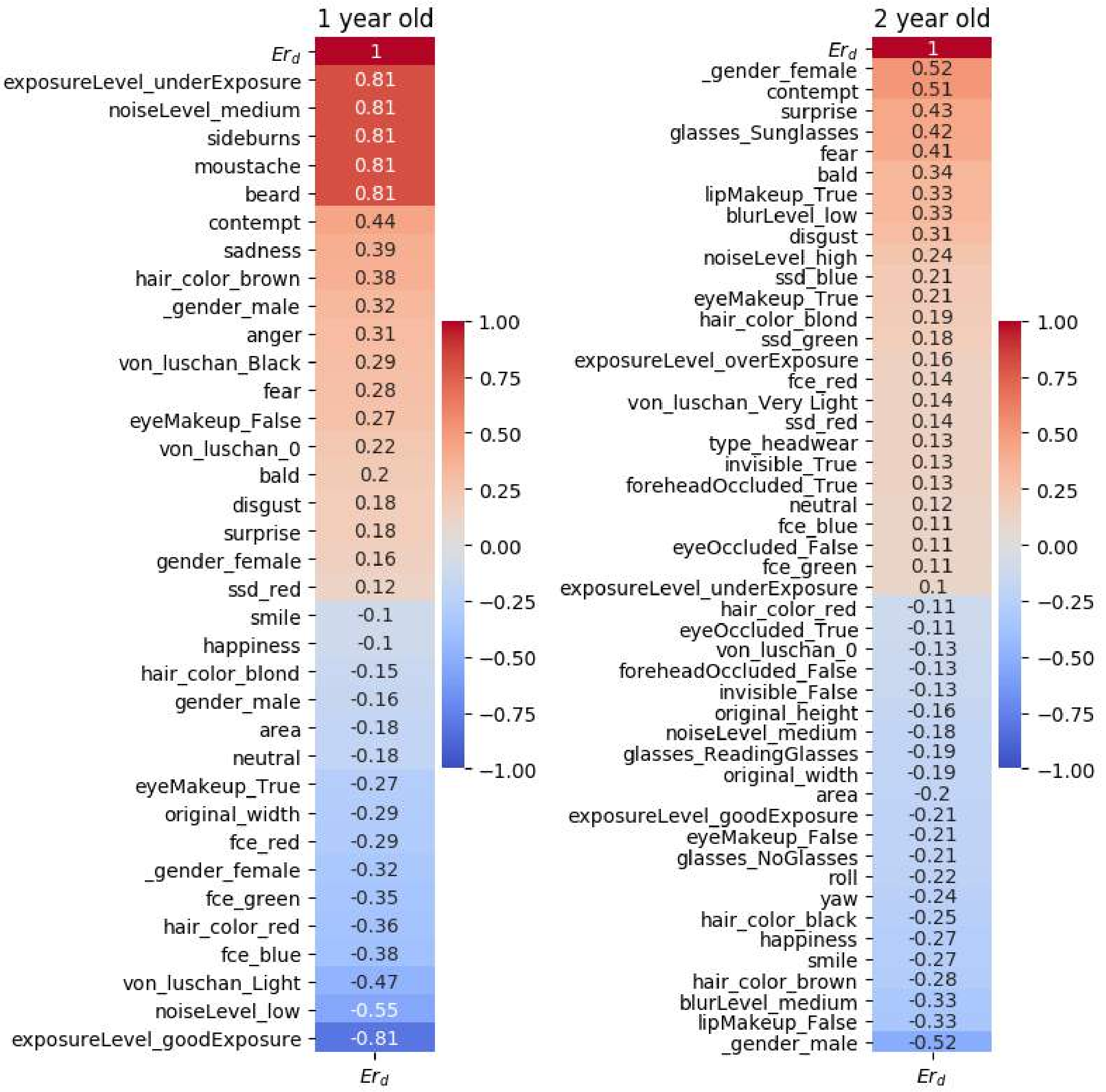


Figure 6.22: Azure: Strong Correlations between Attributes and $\text{Er}_\text{d} > 5$ for Ages 1 and 2.

these ages). In particular, age five has both beard and sideburns attributes with strong correlation PCC values of 0.55. Similarly, both attributes have also been connected to age 6 with correlation PCC values of 0.52 and 0.62 respectively and again on age 10 for beard. Another facial hair attribute, moustache, has also been consistently detected across the 6 to 10 age range as shown on Table 6.11. Overall, it can be deduced that misidentification of facial hair has shown prominence in influencing the decline in the facial age estimator's accuracy for the underage age group. Further research is required to identify the underlying cause of these attributes being detected for the underage age group, particularly under 10s. Conversely, age seven did not present with any correlation towards facial hair. Instead, as shown on Figure 6.23, *blurLevel_high*

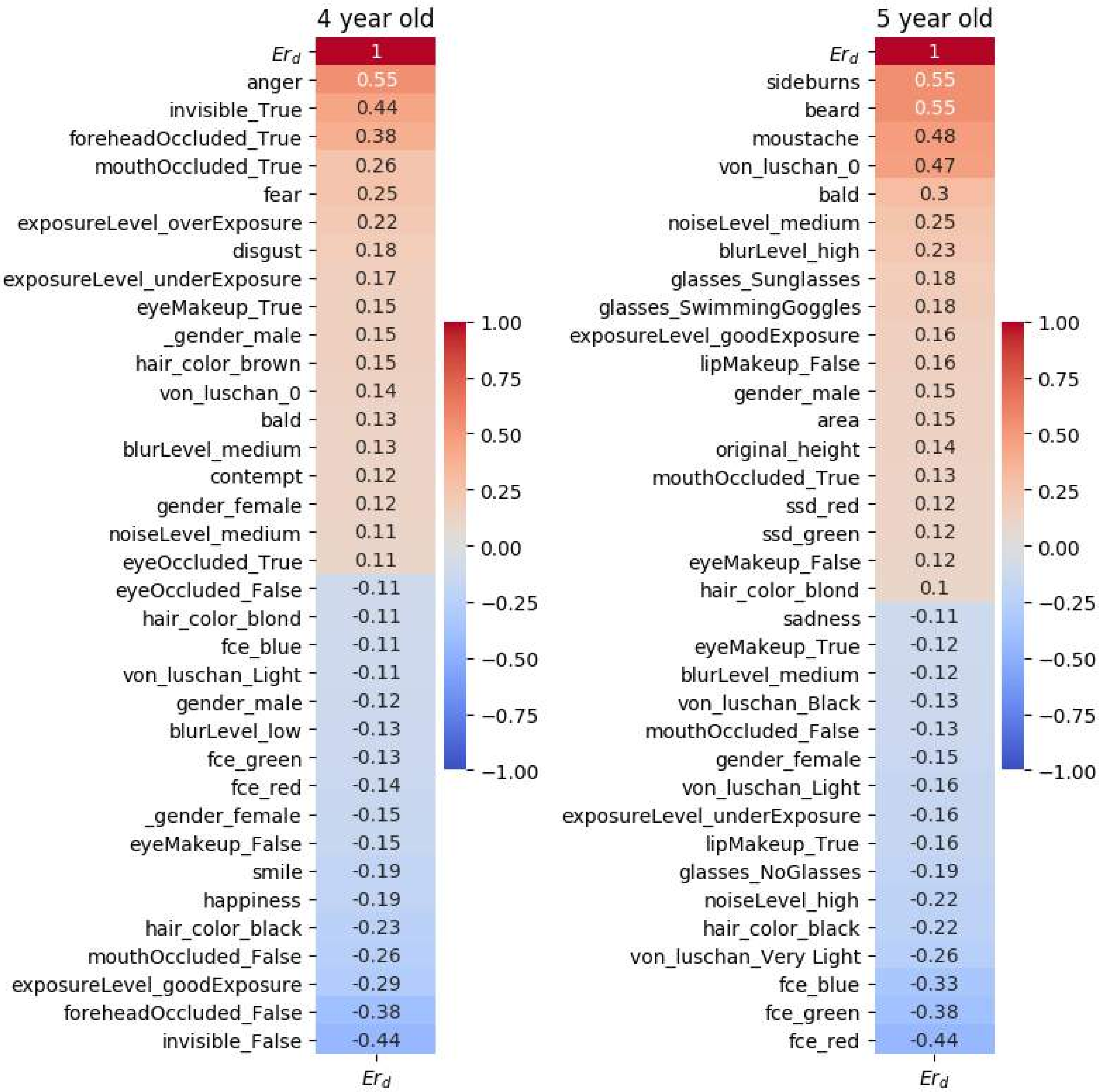


Figure 6.23: Azure: Strong Correlations between Attributes and $\mathrm{Er_d} > 5$ for Ages 4 and 5.

was the only strong correlating attribute detected. Moreover, for age nine, along with the correlation to the moustache facial hair, *SwimmingGoggles* were also found to have strong correlation to the $\mathrm{Er_d}$ with PCC value of 0.78.

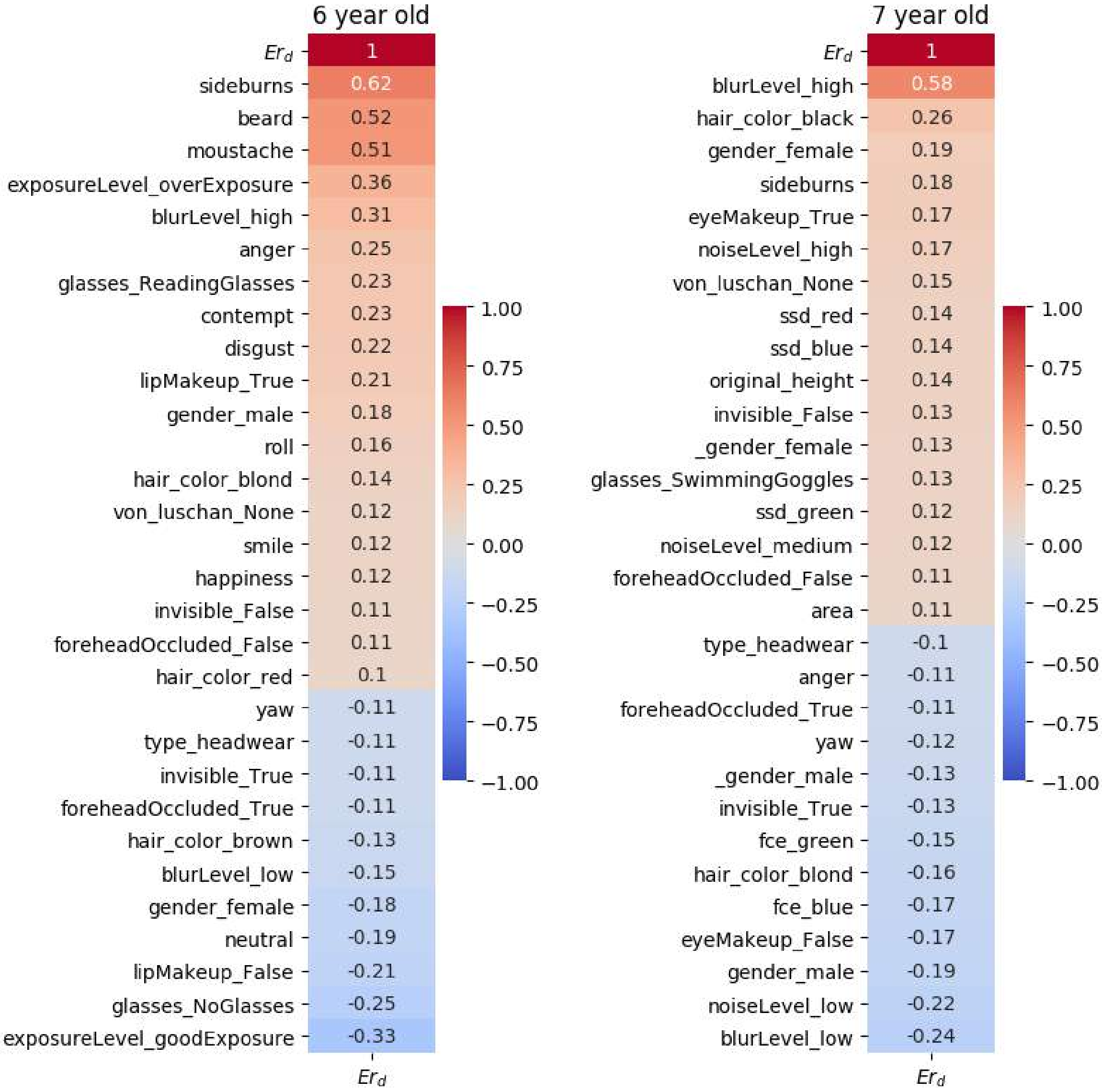


Figure 6.24: Azure: Strong Correlations between Attributes and $\text{Er}_\text{d} > 5$ for Ages 6 and 7.

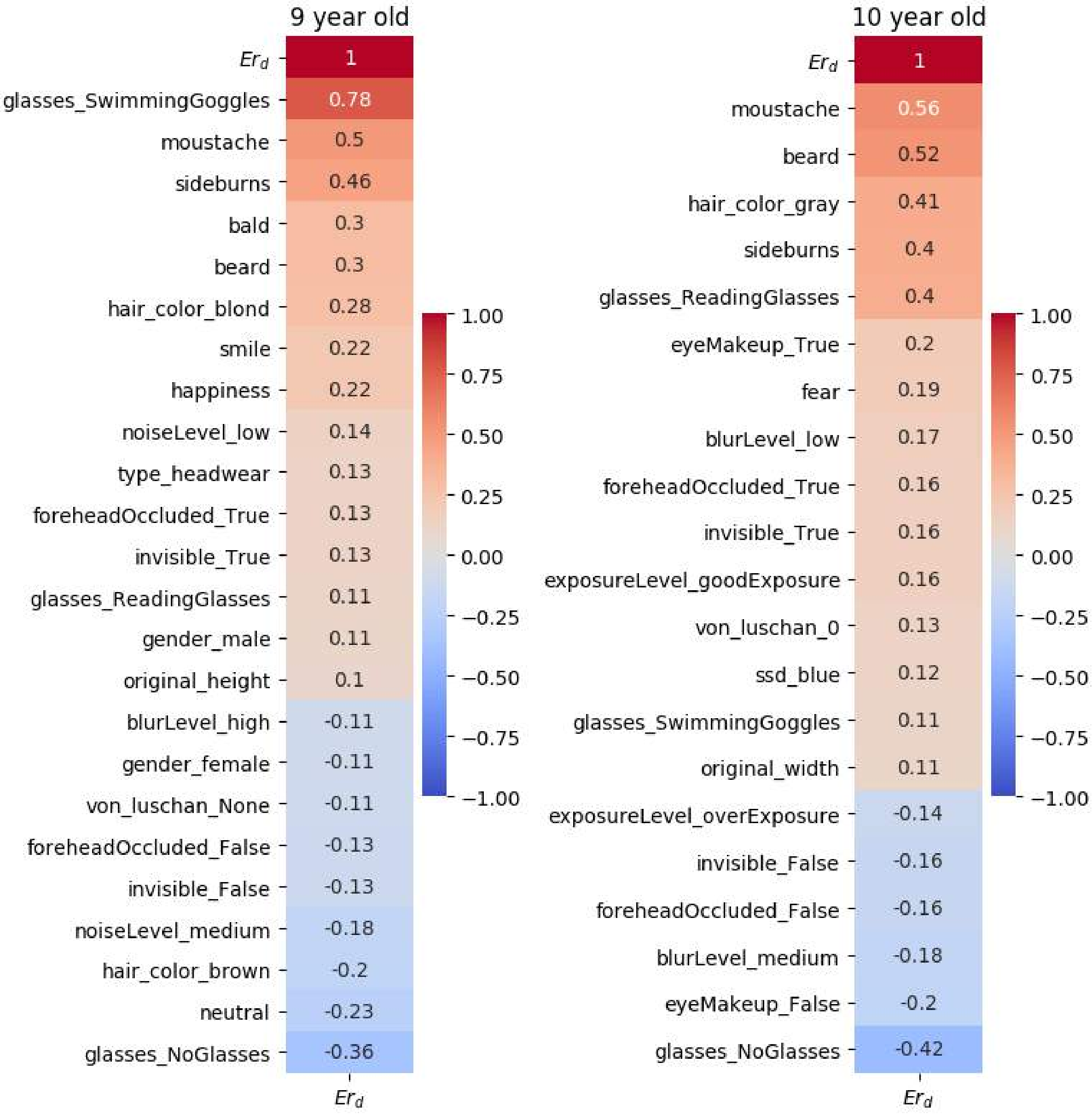


Figure 6.25: Azure: Strong Correlations between Attributes and $\mathrm{Er_d} > 5$ for Ages 9 and 10.

Table 6.11: Azure: Strong (Black) and Mild (Grey) Influencing Factors with $\mathrm{Er_d} > 5$. Note: ages where only weak correlations (between $\mathrm{Er_d}$ and attributes) were discovered are omitted for readability.

| **Degree of Correlation with $\mathrm{Er_d}$** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Attribute Name** | **Age** | | | | | | | | | | |
| | **1** | **2** | **4** | **5** | **6** | **7** | **9** | **10** | **11** | **12** | **17** |
| underExposure | **0.81** | | | | | | | | | | |
| goodExposure | **-0.81** | | | | -0.33 | | | | | | |
| overExposure | | | | | 0.36 | | | | | | |
| noiseLevel_medium | **0.81** | | | | | | | | | | |
| noiseLevel_low | **-0.55** | | | | | | | | | | |
| blurLevel_low | | 0.33 | | | | | | | | | |
| blurLevel_medium | | 0.33 | | | | | | | | | |
| blurLevel_high | | | | | 0.31 | **0.58** | | | | | |
| sideburns | **0.81** | | | **0.55** | **0.62** | 0.46 | | 0.4 | | | |
| moustache | **0.81** | | | 0.48 | **0.51** | | **0.5** | **0.56** | | 0.35 | |
| beard | **0.81** | | | **0.55** | **0.52** | 0.3 | | **0.52** | | | |
| bald | | 0.34 | | 0.3 | | 0.3 | | | | | |
| hair_color_brown | 0.38 | | | | | | | | | | |
| hair_color_gray | | | | | | | | 0.41 | | | |
| hair_color_red | -0.36 | | | | | | | | | | |
| fce_red | | | | -0.44 | | | | | | | |
| fce_blue | -0.38 | | | -0.33 | | | | | | | |
| fce_green | -0.35 | | | -0.38 | | | | | | | |
| gender_female | -0.32 | **0.52** | | | | | | | | | |
| gender_male | 0.32 | **-0.52** | | | | | | | | | |
| contempt | 0.44 | **0.51** | | | | | | | | | |
| anger | 0.31 | | **0.55** | | | | | | | | |
| sadness | 0.39 | | | | | | | | | | |
| fear | | 0.41 | | | | | | | | | |
| disgust | | 0.31 | | | | | | | | | |
| surprise | | 0.43 | | | | | | | | | |
| foreheadOccluded_1 | | | 0.38 | | | | | | | | |
| foreheadOccluded_0 | | | -0.38 | | | | | | | | |
| invisible_True | | | 0.44 | | | | | | | | |
| NoGlasses | | | | | | | -0.36 | -0.42 | -0.32 | | -0.37 |
| SwimmingGoggles | | | | | | | **0.78** | | | | |
| ReadingGlasses | | | | | | | | 0.4 | 0.32 | | 0.37 |
| Sunglasses | | 0.42 | | | | | | | | | |
| lipMakeup_True | | 0.33 | | | | | | | | | |
| lipMakeup_False | | -0.33 | | | | | | | | | |

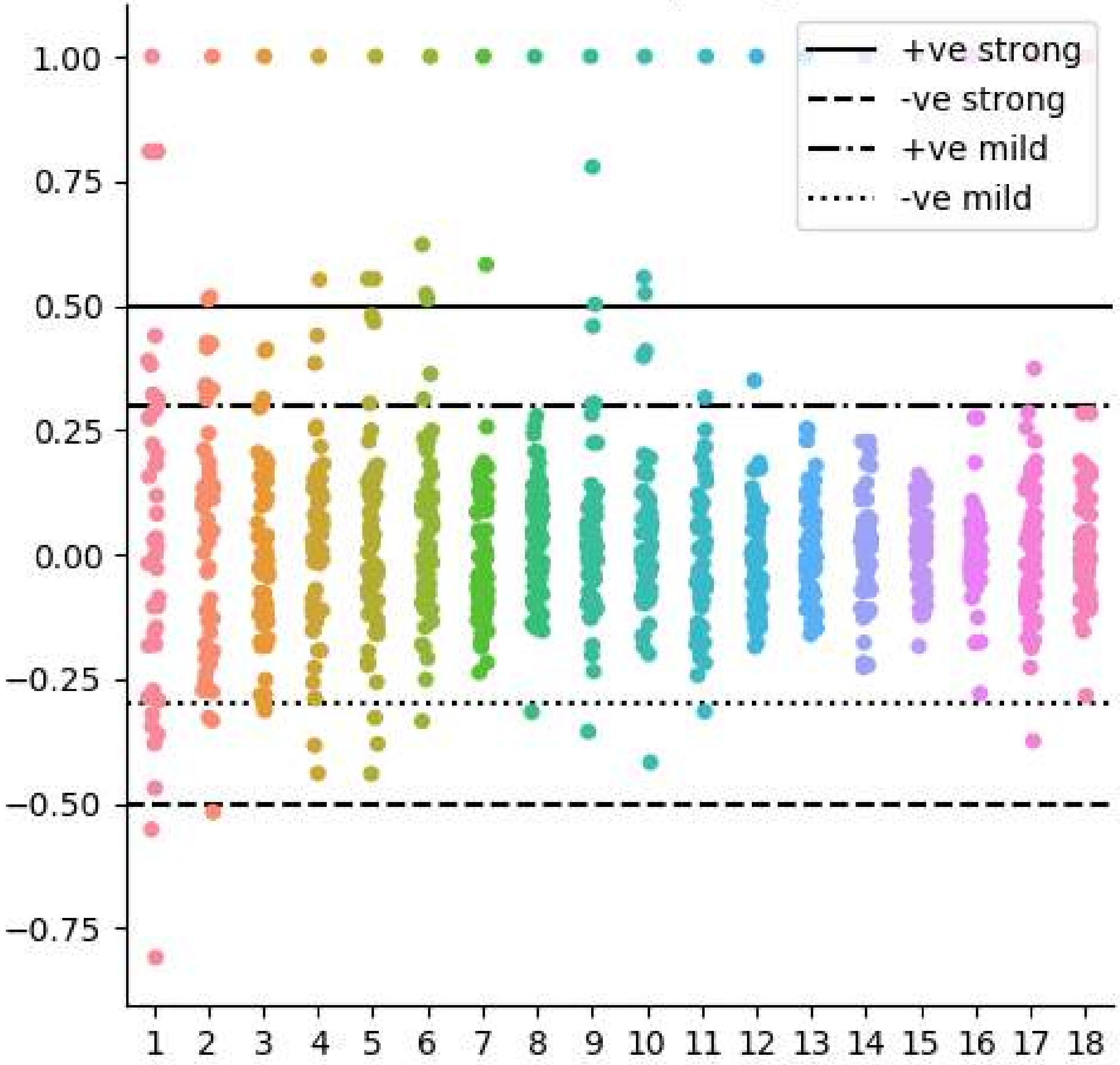


Figure 6.26: Azure: Correlations per Age with $Er_d > 5$.

#### 6.4.1.5 Mild Correlations

Mild correlations have been defined as PCC values between 0.30 to 0.49. Human biometric factors have been playing both strong and mild roles in influencing the accuracy of the age estimations. In addition to the aforementioned biometric attributes, hair colour and skin tone have been found to have mild correlation with $\mathrm{Er_d} > 5$, as shown in Table 6.11. The presence of bald, and brown and grey hair colours on subjects contribute to a higher $\mathrm{Er_d}$. The correlation of *hair_color_gray* with $\mathrm{Er_d}$ was expected as the hair colour is often associated with older adult age ranges. Red hair colour, however, was found to have negative correlation value of -0.36 for one year old's. Furthermore, skin tone (as measured by the FCE attribute) have been detected to have mild correlation to the accuracy of the facial age estimation. In particular, it can be observed that presence of any detected level of FCE on a subject linearly correlates to a more accurate age estimation; *fce_red, fce_blue and fce_green* all demonstrate a negative correlation for ages one and five.

Other bio-metrics that showed strong correlations have also displayed mild correlation values. Mild correlation results for facial hair, as shown on Table 6.11, are inline with the findings in Section 6.4.1.4. Conversely, a bias towards male subjects was highlighted in Section 6.4.1.4. Upon analysing the mild correlations, the female gender attribute has a mild negative PCC of -0.32 verses 0.32 for males. Contempt and anger were the two emotional attributes detected to strongly influence the accuracy of age estimation. In addition to these, emotion of sadness, fear, disgust and surprise were also detected to influence the accuracy of age estimation – however, only in a mild manner. In general, the detection of emotion whether with strong or mild correlation, has linear influence in the decrease of the age estimation performance.

Similarly, the same can be said for the quality of image; the higher level of noise and exposure, presence of blur and occlusion all have linear correlations to higher values of $\mathrm{Er_d}$. Glasses were predominantly found in the older age range – particularly on ages 9, 10, 11 and 17. Its correlation values imply mild to strong correlation with $\mathrm{Er_d}$. Therefore, this is identified as an influencing factor towards Azure's facial age estimation. Moreover, the detection of mild negative correlations of the attribute *NoGlasses* substantiates this finding. This was expected as presence of glasses can distort and provide occlusion to a subjects' face. Similar to glasses, the detection of lip makeup has been found to be mildly associated with $\mathrm{Er_d}$ with attribute *lipMakeup_False* substantiating the result through an opposite correlation with equal gravity.

### 6.4.2 Amazon Rekognition (AWS)

In this section we explore the functionality of AWS, analyse its age estimation accuracy and identify factors that contributed to the results.

#### 6.4.2.1 Error Distribution

Figure 6.27 shows the error tolerance distribution. The majority of errors had low $\mathrm{Er_d}$ (between 0 and 5) signifying that AWS Rekognition's accuracy was within a degree of approximately $\pm 5$ for most underage single-faced images processed. A significantly smaller portion of the age estimations had $\mathrm{Er_d} \geq 10$.

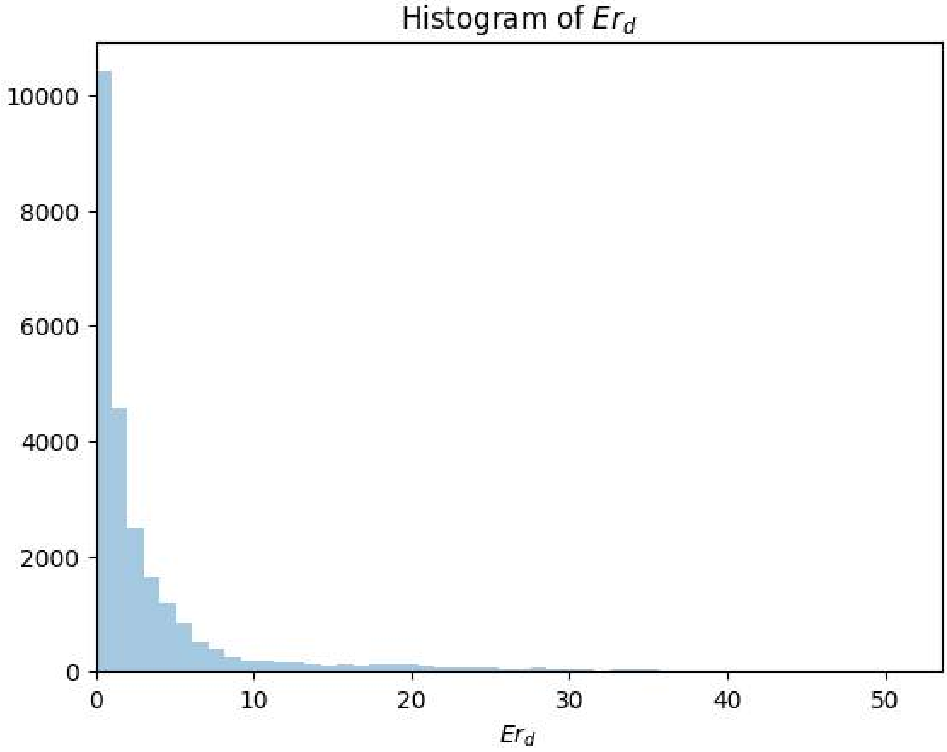


Figure 6.27: AWS: Distribution of $\mathrm{Er_d}$.

#### 6.4.2.2 Strong PCC Distribution per Age with $\mathrm{Er_d}$ greater than 5

Figure 6.28 illustrates the correlations between the attributes and the AWS $\mathrm{Er_d} > 5$. This figure verifies that there are no strong or mild linear correlations between attributes, as shown in Table 4.6. While there are a variety of attributes found to have weak associations with $\mathrm{Er_d} > 5$, there are no strong influencing factors that affect the AWS

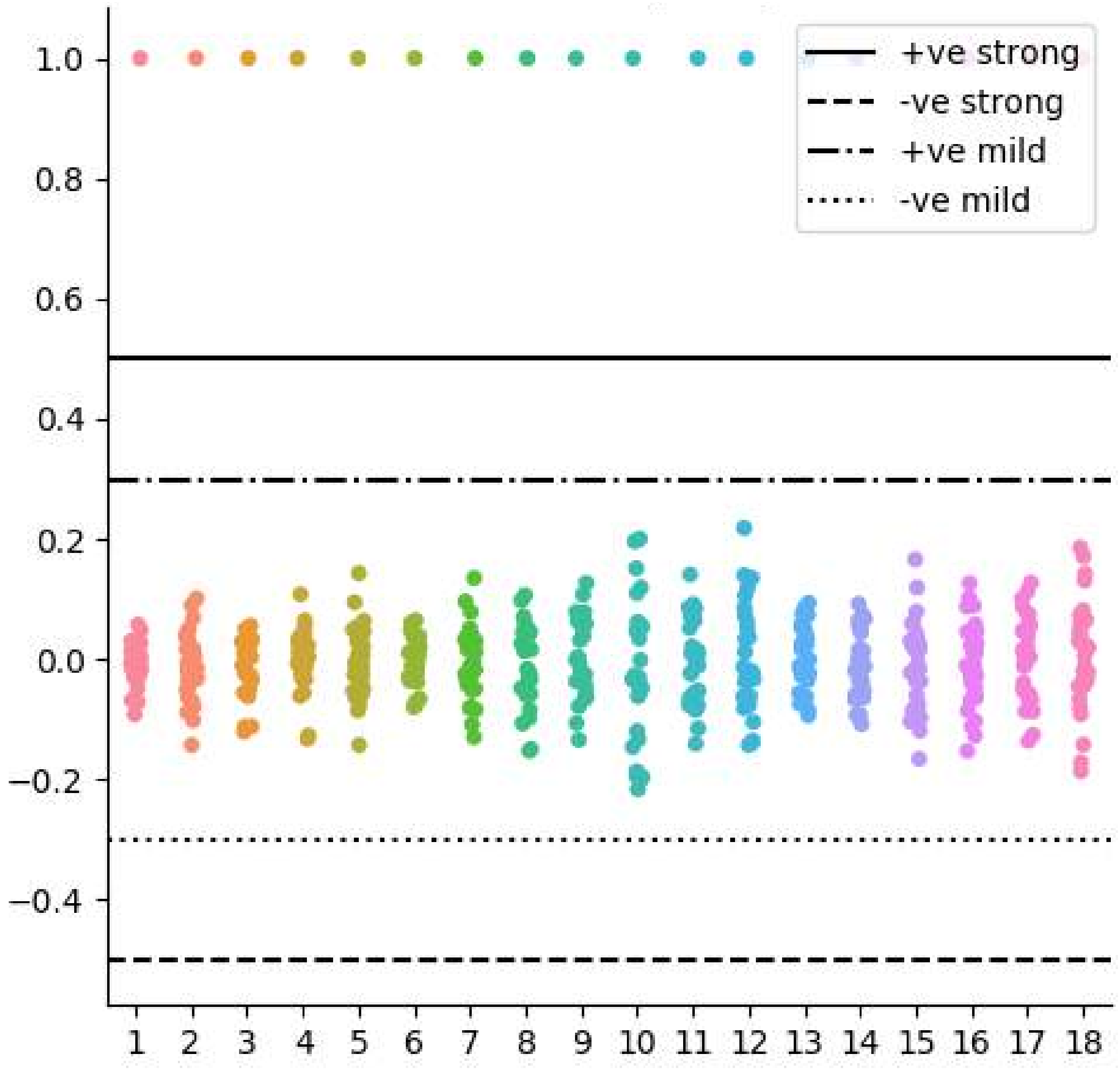


Figure 6.28: AWS: Correlations per Age with $\mathrm{Er_d} > 5$.

accuracy when the error margin is greater than 5, as shown in Table 4.6. This investigation was replicated for $\mathrm{Er_d} \geqslant 0$ and $\mathrm{Er_d} \leqslant 5$ inline with the investigation process used for Azure. A similar result with $\mathrm{Er_d} > 5$ obtained for all other values of $\mathrm{Er_d}$. There were no strong correlations identified. Therefore, from the conclusive results obtained for AWS Rekognition, it can be concluded that there are no influencing factors that contribute to the magnitude of its facial age estimation accuracy. Baring in mind that these correlation results are based on PCC, mild to strong nonlinear correlation may still exist. Further study is required to investigate potential nonlinear correlations.

## 6.5 DeepUAge

An age balanced dataset of 16,000 images was prepared. This dataset was discussed in Section 4.5.2.1. 80% of the images were used for training and 20% were used for validation. The training stopped on 87 epochs, maintained a loss under 1.799 with a favourable MAE of 1.57 years and 2.79 years for training and validation respectively. The model was further tested with 1,000 additional images that were gathered from the UTKFace dataset [270] and the dataset generator. These images amounted to 50 images per class. The testing achieved a MAE of 2.73 years.

In the process of creating the DeepUAge model, the DCA approach was developed. Its efficiency in pre-processing has been evaluated in comparison to other pre-processing techniques. It produced a MAE of 2.73 years, which was the best performing of all the approaches evaluated. The results for this experiment can be found in Table 6.12. The accuracy of the DeepUAge model in both validation and testing in comparison with other facial age estimators for underage subjects is outlined in Section 6.5.1.3.

| Approach | MAE |
|---|---|
| **DCA** | **2.73** |
| Face++ contour | 2.79 |
| `Dlib` contour aligned | 4.01 |
| `Dlib` contour non-aligned | 4.28 |
| `Dlib` cropped | 5.31 |
| Non-processed | 5.71 |

Table 6.12: Results of Different Pre-Processing Techniques

### 6.5.1 Evaluation

#### 6.5.1.1 DCA and Other Pre-Processing Techniques

By separating the age estimation problem into a smaller scope of age range, it was possible to validate the input data of juveniles used to train the model. As the data used was not only frontal “passport photo style” facial images, pre-processing procedures have been adopted to minimise any negative impact of of noise on the final results.

Several types of existing pre-processing techniques were evaluated; `dlib` contour aligned, `dlib` contour non-aligned, `dlib` cropped, and Face++ contour. Variations of the `dlib` approaches all produced a MAE greater than 4 years. In particular, the `dlib` cropped technique was found to perform almost equally as not using a pre-processing filter at all. Additionally, even the best performing `dlib` contour aligned, was still 43.72% greater in MAE versus the Face++ contour. But, due to the unsuitability of Face++ for CSEM investigation, the technique was deemed unusable for the DeepUAge model. This motivated the development of the DCA facial cropping techniques, which nonetheless achieved better results than all other pre-processing techniques evaluated.

#### 6.5.1.2 DeepUAge Performance

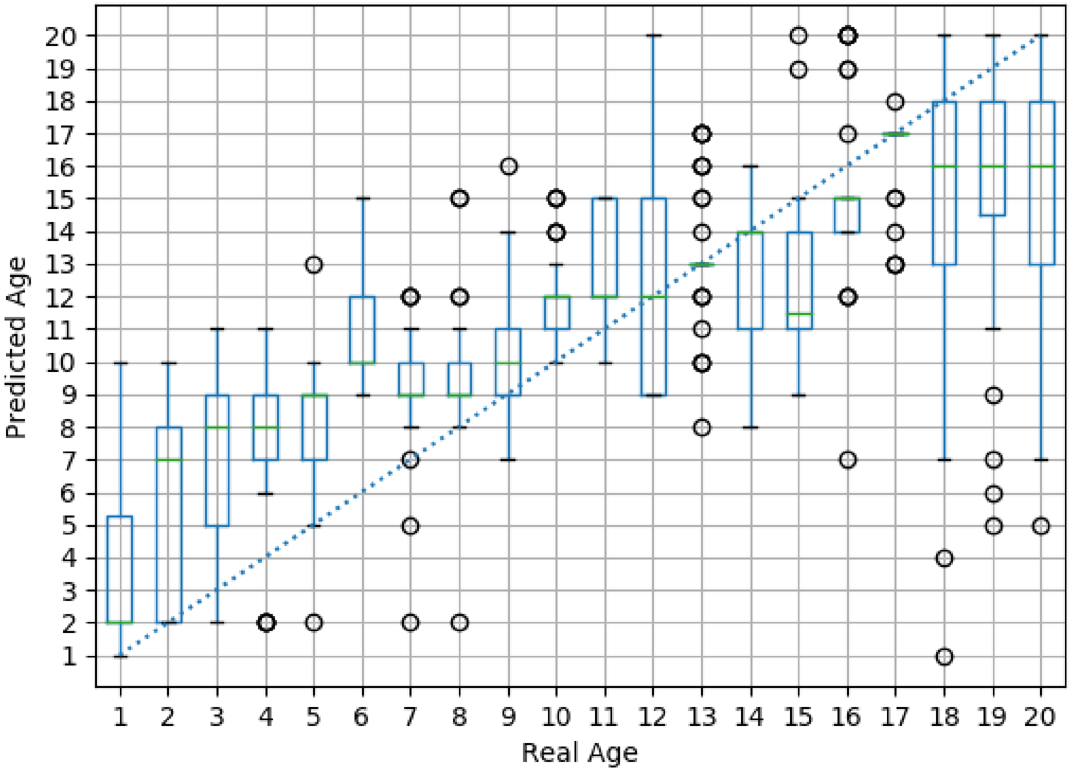


Figure 6.29: DeepUAge - Real Age vs Predicted Age

Figure 6.29 illustrates a box plot of real age vs. DeepUAge predicted age where accuracy and precision for the age range 1-20 can be observed. Ages 18 and 20 have the largest range of predicted age. Whilst the model manages to obtain correct classification for both ages, the performance score for these classification age ranges is low. Instead, the correct age prediction is found at the maximum whiskers of the data for the age of 20 and at the 3rd quartile of the data distribution for the age of 18.

Similar results can be seen across the age range 14 to 20 where the correct age lies between the maximum and 3rd quartile of the classification distribution. Conversely, the reverse can be observed on the opposite end of the age spectrum where younger ages are often overestimated. It can be observed that for ages 1, 5, 8 and 10 the minimum of the classification distribution contained the correct catalogue of the images. Furthermore, for ages 2, 3, 9 and 11 the correct predicted age is between the minimum and the 1st quartile of the classification distribution. There are several outliers across all ages, predominately in ages 7, 13, 16, 17 and 19, where 4 or more outliers can be observed.

#### 6.5.1.3 Comparison of DeepUAge and Other Estimation Techniques

State-of-the-art age estimators were compared against DeepUAge, with the same testing set used for the pre-processing comparisons, and the same pre-processing approach (DCA approach applied to DeepUAge). The test performance achieved is inline with the state-of-the art age estimation classifiers. The testing MAE of 2.73 for DeepUAge indicates that the average magnitude of the model error was significantly lower than that of its alternatives. The top 3 performers found in descending order were DeepUAge, Microsoft Azure Face API, followed by Amazon Rekognition, as can be seen in Table 6.13. Both Microsoft's and Amazon's approaches were side by side in performance with only being 0.14 MAE apart. DeepUAge stood at 0.87 MAE better than Microsoft Azure Face API and exceeding Amazon Rekognition by a MAE of 1.01.

| Approach | MAE |
|---|---|
| **DeepUAge Test** | **2.73** |
| Microsoft Azure Face API | 3.60 |
| Amazon Rekognition | 3.74 |
| Dummy estimator (All assumed 10 y/o) | 5.00 |
| Face++ | 18.21 |
| IMDB-WIKI WideResNet [218] | 20.43 |

Table 6.13: Evaluation of Facial Age Estimators for Underage Subjects

Despite obtaining the best results of the age estimators evaluated, DeepUAge has a logarithmic loss (the variation of the actual label from the machine learning model predicted value) of 1.799. Our goal is to decrease this value further to as close to 0 as possible and is addressed as future work. Although IMDB-WIKI WideResNet was trained on a large celebrity dataset (over half a million images), it had the lowest performance. Along with Face++, both estimators had the poorest performance overall. The dummy estimator (an approach which classifies images to a fixed predicted age of 10) managed to surpass the performance of two intricate algorithms. This result can be due to the failure to validate age labels and the lack of images in the underage age group.

As shown in Table 6.13, the overall performance of DeepUAge was found to be best compared with the other models evaluated. This is demonstrated at a better granularity in Figure 6.30 where DeepUAge is shown to catalogue age 9 to 18 with a higher accuracy than those of the other age predicting services. In reference to Figure 6.29, it can be further concluded that the precision of DeepUAge is at its peak for ages 7, 8, 10, 13, 16 and 17. In particular, it is most accurate at age classification of 13 and 17 year old's; therefore providing better age classification techniques for "early adolescent (age 10 to 14) and late adolescent (age 15 to 19) years" [231]. This age range is important in CSEM investigations because it covers subjects within the borderline of adulthood.

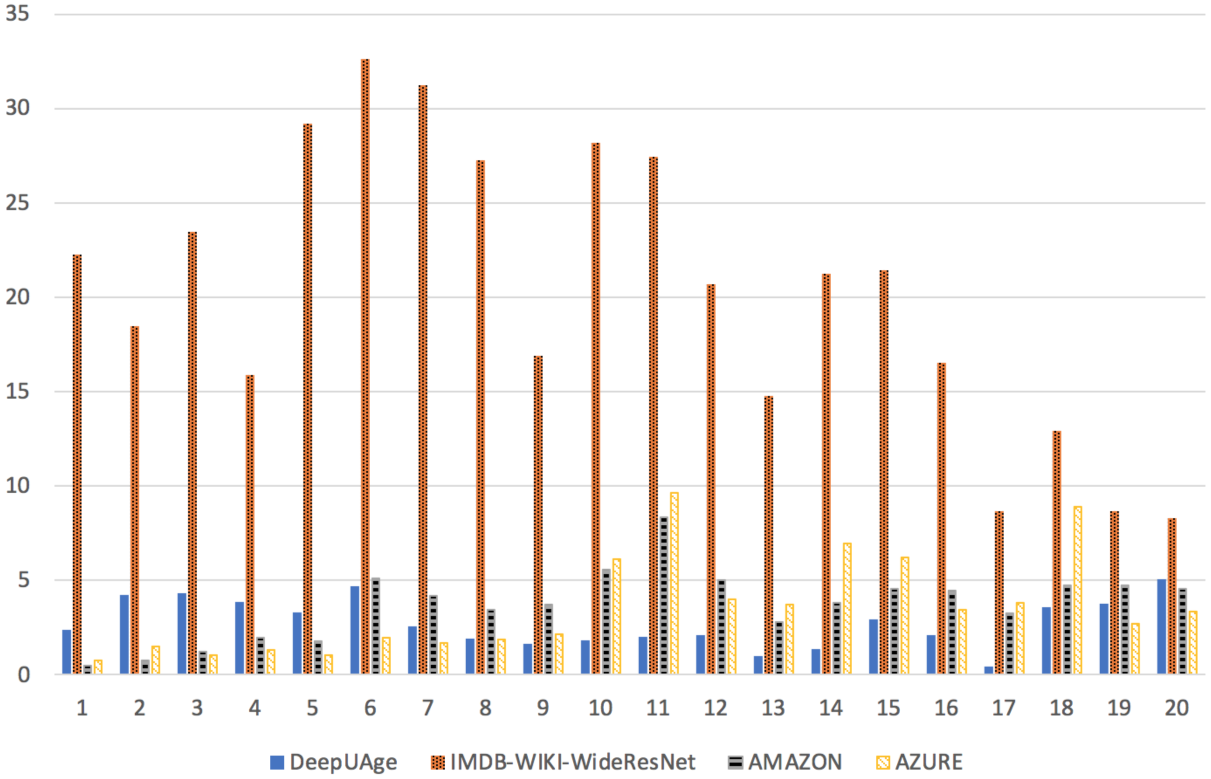


Figure 6.30: Average Difference per Age per Facial Age Predictor.

The performance of online age estimation services (for the underage group) such as AWS and Azure without the DCA approach can be observed in Table 6.4. With a MAE of 3.349 and 5.347 respectively. Although the dataset employed is quite similar, it can be observed that with the DCA approach there was an improvement only for the Microsoft Azure Face API approach.

## 6.6 Vec2UAge

The mixed dataset discussed in Section 4.5.3.1 was employed. All faces in the images were detected and cropped, as described in Section 4.5.3.2, and the facial vectors calculated. Images for training were generated with the relevant data augmentation techniques, as explained in Section 4.5.3.4, to a total of 90k (5k per class, 18 classes in total), a stratified shuffle split was applied to the test dataset to divide it into 2 equal sub-datasets. Both the validation and testing dataset accounted to 4,500 images each.

A 4-layered neural network was selected with 512, 256, 128 and 1 units per layer respectively. Each hidden layer used a ReLU activation function. The input to the network was the array of 512 facial vectors and the output an age regressor. The optimisation algorithms chosen were ADAM, ADAGRAD, SGD and SWA. Two sets of 20 experiments each were performed with the aid of Neptune, a light-weight management tool that keeps track of ML experiments [186]. The first set of experiments, *D1*, correspond to a initial fixed LR of `1e−4` for ADAM, ADAGRAD and SGD, and `1e−5` for SWA. Conversely, the second set of 20 experiments (E1) correspond to the use of a tool as guidance for choosing an optimal initial LR. This tool is based on cyclic learning rates proposed by Smith [233]

The choice of the loss function was the simplest and most common MSE. The main metrics used to measure the loss were MSE and MAE. The MAE is the absolute mean average difference between the predicted age and the real age. Lastly, the number of epochs selected was 100, but early stopping was implemented. This is helpful to reduce the LR as the number of training epochs increases and therefore, a LR scheduler was applied.

### 6.6.1 Evaluation of the Experiments

The set of experiments *D1* and *E1* have been logged entirely using Neptune. Up to 10 experiments can be compared simultaneously and there is an API feature that allows the integration with python through the *neptune.sessions* library. To ensure full reproducibility from run to run, *pytorch-lightning* supports deterministic experiments. Additionally, the seeds for pseudo-random generators have been logged in Neptune and the experiment results are duplicable and openly available at `https://ui.neptune.ai/4nd4/Vec2UAge/`.

### 6.6.2 D1 - Evaluation of the MAE Distribution with a Fixed Initial Learning Rate

The effects of the fixed initial values of *LR* (`1e−4` for ADAGRAD, ADAM and SGD, and `1e−5` for SWA) can be seen in Figure 6.31. *ADAGRAD* yielded the worst performing results for training, validation and test. Nevertheless, the consistency of values of the validation and test loss is denoted by a spread of 0.05 and visible in both the figure and in Table 6.14. It can be seen that the *ADAM* algorithm surpassed the performance of the other optimisers for training and testing. Moreover, the data was the least sparse for all the losses. The validation and test MAE for *SGD* was consistent and the standard deviation was low. Therefore, there was not a significant spread of data. It was a stable optimiser that produced models as low as 2.51. Lastly, *SWA* although not achieving the best performing training and validation values, managed to achieve models for validation as low as 2.43, and had the best performing model and mean for testing.

In terms of the difference between training and test/validation sets, for ADAGRAD, the mean training accuracy is the same as the mean testing accuracy but testing has the data less disperse. For ADAM, the mean training MAE is the lowest, validation and test have the same mean and variance. For SGD, the mean training accuracy is lower than both validation and test, but test has the lowest value. Finally, for SWA, the testing accuracy is better than the validation accuracy and has less variance. It would usually be expected that the training accuracy surpasses validation and testing. Also it is quite common that the testing accuracy is lower than the validation accuracy. But in these set of experiments, there are certain behaviours that could have happened due to data bias.

Overall, the best performer for the experiment set *D1* was the *ADAM* optimiser approach with a fixed initial LR of `1e−4`. The outcome produced models in test with a mean of 2.49 and MAEs as low as 2.46. The criteria to pick the best optimiser was to sum the count of minimum values per statistic per loss.

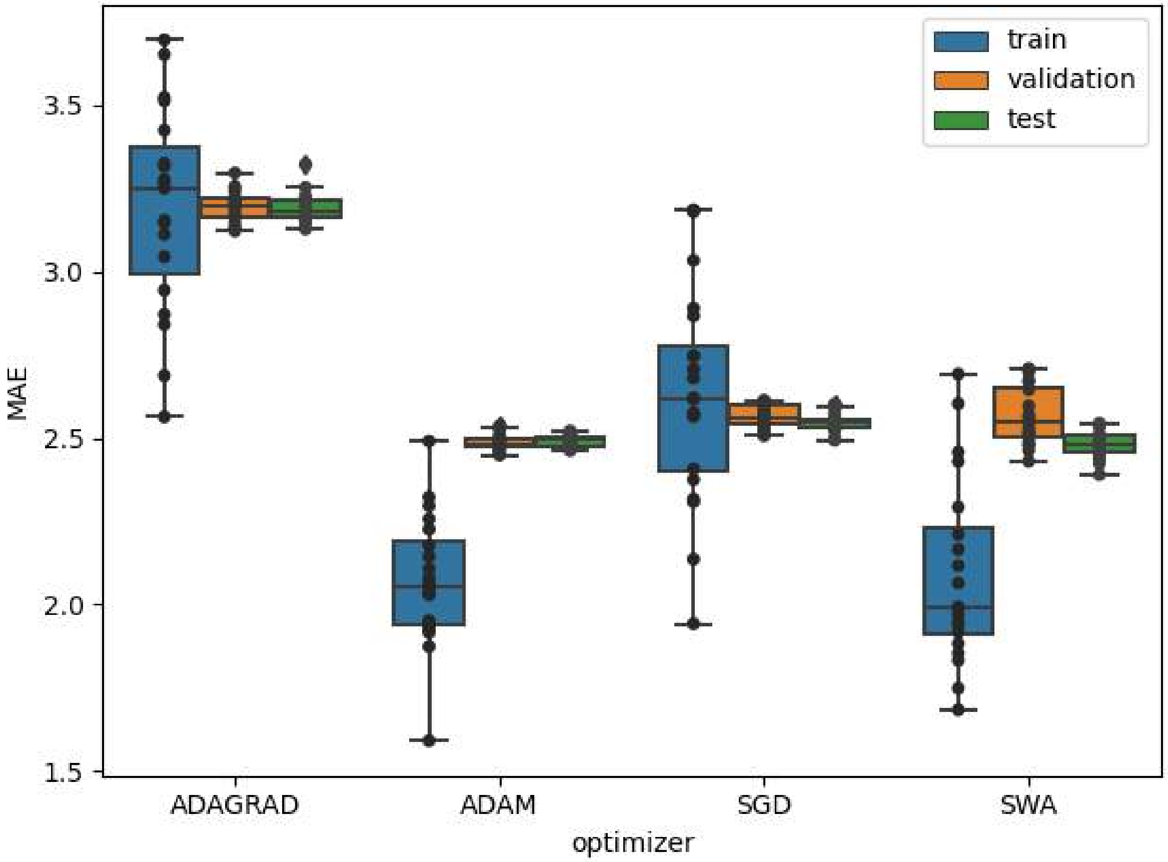


Figure 6.31: Mean Absolute Error Distribution per Optimiser for Training/Validation/Test (Using a Fixed Initial Learning Rate)

| | **mae** | | | **val_mae** | | | **test_mae** | | |
|---|---|---|---|---|---|---|---|---|---|
| **optimiser** | *min* | *mean* | *std* | *min* | *mean* | *std* | *min* | *mean* | *std* |
| *ADAGRAD* | 2.56 | 3.19 | 0.31 | 3.12 | 3.20 | 0.05 | 3.13 | 3.19 | 0.05 |
| *ADAM* | 1.59 | 2.07 | 0.20 | 2.45 | 2.49 | 0.02 | 2.46 | 2.49 | 0.02 |
| *SGD* | 1.94 | 2.62 | 0.32 | 2.51 | 2.56 | 0.03 | 2.49 | 2.55 | 0.03 |
| *SWA* | 1.68 | 2.09 | 0.28 | 2.43 | 2.57 | 0.09 | 2.39 | 2.48 | 0.04 |

Table 6.14: Experiment D1: Statistics of Mean Absolute Error in Training/Validation/Test (Fixed Learning Rate of `1e−4` for ADAGRAD, ADAM and SGD; and `1e−5` for SWA)

### 6.6.3 E1 - Evaluation of the MAE Distribution with an Initial LR Finder

The effects of the usage of the LR finder to obtain the initial learning rate can be seen in Figure 6.32. It is observed that the validation values are consistently clustered except for the ones seen in the ADAM algorithm. *ADAGRAD* had low sparse data throughout training, validation and test losses. Conversely, the performance was not as good as the rest of the optimisers (ADAM, SGD, SWA). It is noticeable that *ADAM* had the highest standard deviation figures with 1.23, 0.93 and 0.92 in training, validation and testing, respectively, as can be seen in Table 6.15. Next, *SGD* performed better than the rest of the optimisers followed by *SWA*. The winning results were consistent for all the statistical evaluations besides the standard deviation for training, which was slightly inferior than *SWA*. Experiment E1 produced models with MAEs as low as 2.36. The same criteria used in Section 6.6.2 to pick the best optimisers was applied.

In terms of the difference between training and test/validation sets, for ADAGRAD, the mean accuracy behaves as expected. Although the test accuracy is close to the validation accuracy, the validation has less dispersion. For ADAM, the mean accuracy for test is higher than validation but both have high standard deviation values. For SGD, the mean training accuracy is lower than both validation and test which have the same mean values. Finally, for SWA, the testing accuracy is better than the validation accuracy and has less variance. In general, the values from experiment *E1* are more accurate with an exception of the ADAM optimiser. Nevertheless the variation in training/testing/validation is in average higher for *E1*.

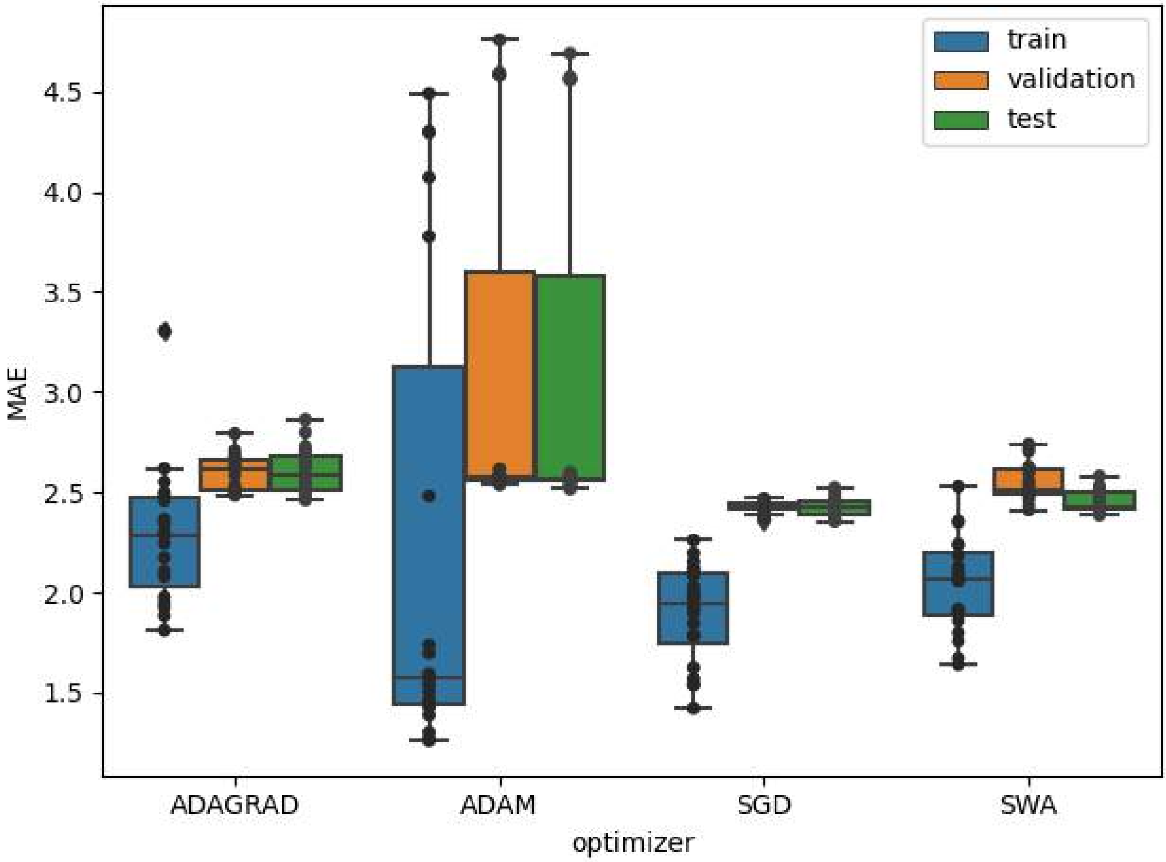


Figure 6.32: Mean Absolute Error Distribution per Optimiser for Training/Validation/Test (Using LR Finder)

| | **mae** | | | **val_mae** | | | **test_mae** | | |
|---|---|---|---|---|---|---|---|---|---|
| **optimiser** | *min* | *mean* | *std* | *min* | *mean* | *std* | *min* | *mean* | *std* |
| *ADAGRAD* | 1.81 | 2.28 | 0.35 | 2.48 | 2.60 | 0.09 | 2.46 | 2.61 | 0.12 |
| *ADAM* | 1.26 | 2.24 | 1.23 | 2.54 | 3.11 | 0.93 | 2.52 | 3.09 | 0.92 |
| *SGD* | 1.42 | 1.89 | 0.25 | 2.36 | 2.43 | 0.03 | 2.36 | 2.43 | 0.05 |
| *SWA* | 1.64 | 2.04 | 0.24 | 2.41 | 2.55 | 0.09 | 2.38 | 2.46 | 0.05 |

Table 6.15: Experiment E1: Statistics of Mean Absolute Error in Training/Validation/Test (LR Finder Executed)

### 6.6.4 Details of Winning Optimisation Approaches

The best performer in *D1* is the experiment *VEC-403* with a validation MAE of 2.48 and a test MAE of 2.39. The next best performer is *VEC-394* with values of 2.51 and 2.42 for validation and test losses respectively. The numbers of each experiment with the corresponding seed and losses can be seen in Table 6.16.

|  | id | seed | val_mae | test_mae |
|---|---|---|---|---|
| 1 | *VEC-382* | 7399 | 2.43 | 2.50 |
| 2 | *VEC-383* | 2125 | 2.71 | 2.48 |
| 3 | *VEC-384* | 7889 | 2.51 | 2.46 |
| 4 | *VEC-386* | 1167 | 2.67 | 2.48 |
| 5 | *VEC-387* | 3512 | 2.52 | 2.54 |
| 6 | *VEC-388* | 9970 | 2.65 | 2.47 |
| 7 | *VEC-389* | 7698 | 2.71 | 2.45 |
| 8 | *VEC-391* | 8659 | 2.53 | 2.50 |
| 9 | *VEC-392* | 8010 | 2.57 | 2.54 |
| 10 | *VEC-393* | 6904 | 2.49 | 2.47 |
| 11 | *VEC-394* | 4422 | 2.51 | 2.42 |
| 12 | *VEC-396* | 7310 | 2.48 | 2.44 |
| 13 | *VEC-397* | 2086 | 2.60 | 2.44 |
| 14 | *VEC-398* | 6547 | 2.56 | 2.51 |
| 15 | *VEC-399* | 3781 | 2.67 | 2.53 |
| 16 | *VEC-401* | 6587 | 2.70 | 2.47 |
| 17 | *VEC-402* | 9677 | 2.46 | 2.51 |
| 18 | *VEC-403* | 6569 | 2.48 | 2.39 |
| 19 | *VEC-404* | 3131 | 2.56 | 2.49 |
| 20 | *VEC-406* | 75 | 2.55 | 2.51 |

Table 6.16: Experiment D1: Validation Mean Absolute Error with Stochastic Weight Averaging Optimiser and Fixed LR of `1e−5`. The top 3 performers are highlighted.

The best performer in *E1* is the experiment *VEC-295* with a validation MAE of 2.44 and a test MAE of 2.36. The next best performer is *VEC-286* with values of 2.46 and 2.36 for validation and test respectively. Both leading experiments had almost the same outcome; the numbers of each experiment with the corresponding seed, *LR* and losses can be seen in Table 6.17.

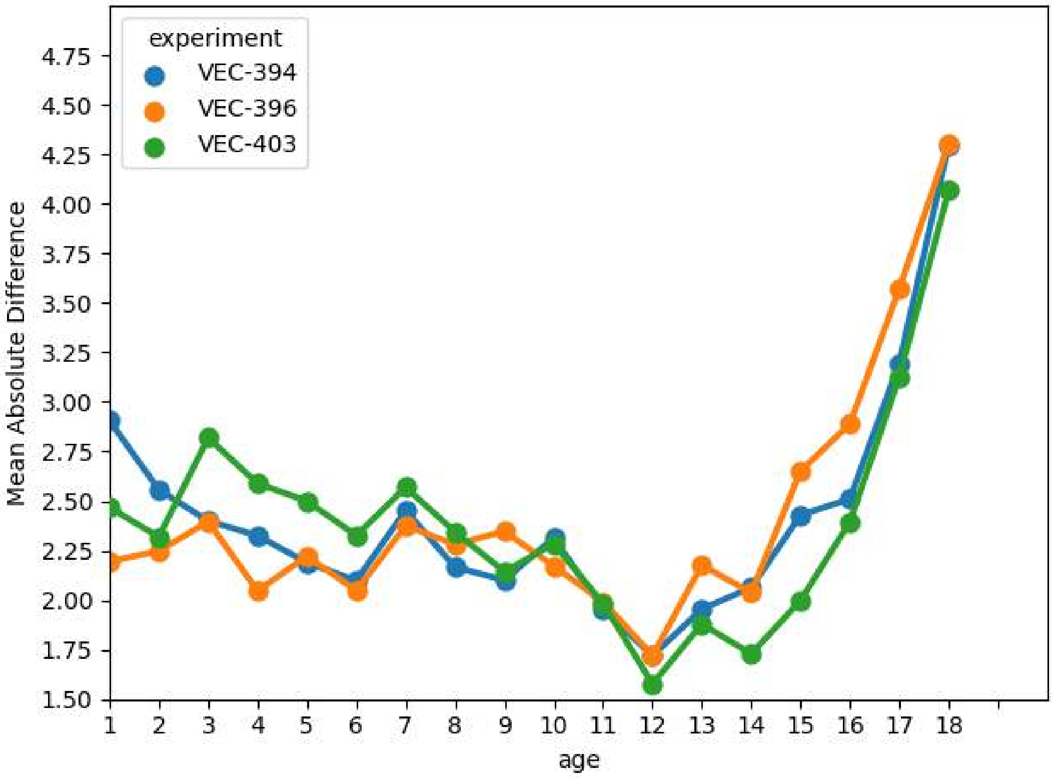
The winning model (experiment *VEC-295*) was an outcome of a SGD optimisation approach with an initial optimal LR of 0.0302. The model produced a MAE in validation and test of 2.46 and 2.36 respectively. The performance per age for the top 3 best performing experiments *D1* and *E1* can be seen in Figure 6.33. The winning model had a MAD performance range between 1.84 and 4.47. The performance was at its best for 2, 12 and 14 year-old subjects. The trend of the other experiments (*VEC-295, VEC-286 & VEC-311*) are similar. As can be seen in Figure 6.33b, they each perform well for 12-year-old subjects and the performance starts decreasing from 14 year-old's on-wards in an exponential manner. In a similar way, the top 3 best performers for experiment *D1* have a behaviour inline with experiment *E1* with a good performance in 12 and 14 year-old subjects while having an exponential decrease in performance from 14 year-

|  | id | seed | lr | val_mae | test_mae |
|---|---|---|---|---|---|
| 1 | *VEC-283* | 6628 | 0.0251 | 2.36 | 2.52 |
| 2 | *VEC-286* | 7122 | 0.0209 | 2.46 | 2.36 |
| 3 | *VEC-291* | 426 | 0.0251 | 2.46 | 2.45 |
| 4 | *VEC-295* | 6532 | 0.0302 | 2.44 | 2.36 |
| 5 | *VEC-301* | 420 | 0.0209 | 2.47 | 2.37 |
| 6 | *VEC-304* | 9010 | 0.0251 | 2.44 | 2.42 |
| 7 | *VEC-307* | 958 | 0.0251 | 2.39 | 2.44 |
| 8 | *VEC-311* | 6795 | 0.0251 | 2.44 | 2.38 |
| 9 | *VEC-314* | 1298 | 0.0251 | 2.45 | 2.41 |
| 10 | *VEC-321* | 9701 | 0.0363 | 2.39 | 2.40 |
| 11 | *VEC-325* | 8329 | 0.0251 | 2.44 | 2.42 |
| 12 | *VEC-330* | 4202 | 0.0251 | 2.42 | 2.46 |
| 13 | *VEC-334* | 1532 | 0.0302 | 2.41 | 2.42 |
| 14 | *VEC-337* | 2646 | 0.0251 | 2.43 | 2.39 |
| 15 | *VEC-340* | 2952 | 0.0251 | 2.42 | 2.46 |
| 16 | *VEC-343* | 2158 | 0.0437 | 2.44 | 2.40 |
| 17 | *VEC-346* | 5482 | 0.0251 | 2.41 | 2.44 |
| 18 | *VEC-349* | 4667 | 0.0251 | 2.44 | 2.44 |
| 19 | *VEC-352* | 6527 | 0.0251 | 2.37 | 2.48 |
| 20 | *VEC-354* | 6811 | 0.0251 | 2.42 | 2.50 |

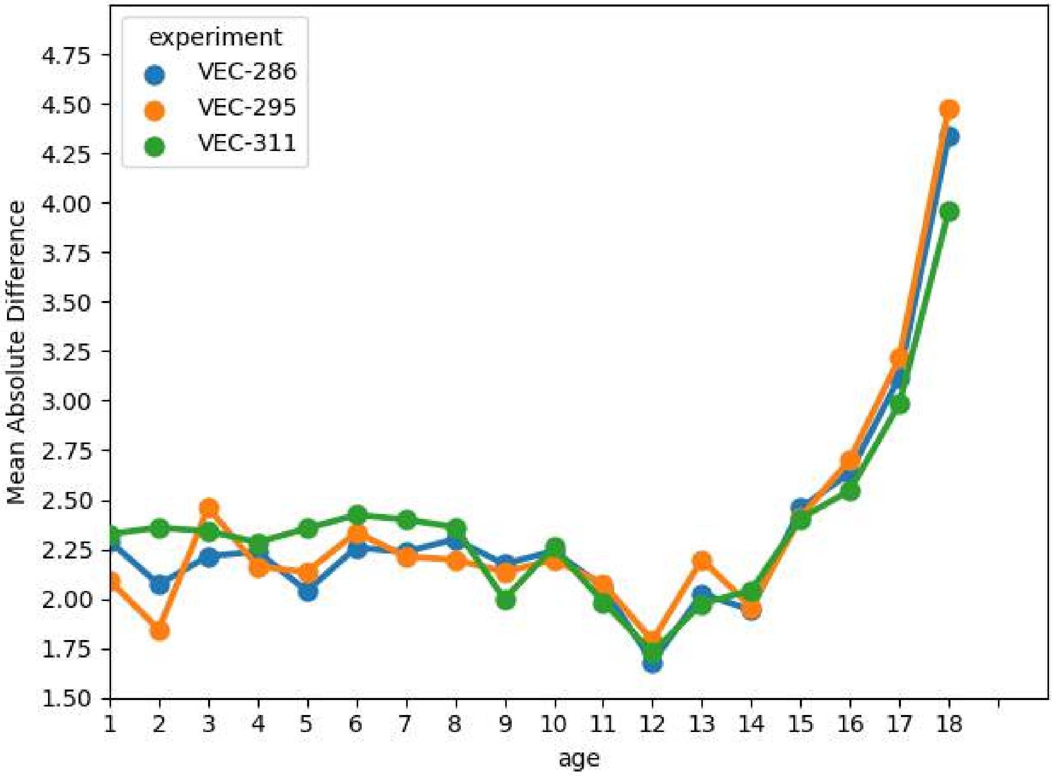
Table 6.17: Experiment E1: Validation Mean Absolute Error with Stochastic Gradient Descent Optimiser and LR Finder [233]. The top 3 performers are highlighted.

old's on-wards, as can be seen in Figure 6.33a.

In both experiments, it is observed that in the age range 14 to 18, there are significant changes of MAD. This age range coincides with the teen group (with 14 being in some cases the beginning of puberty) where the facial vector approach starts decreasing performance. However, our model performs well under a MAE of 2.5 (which surpasses state-of-the-art performance) for 1 year-old's to preteens.



(a) Experiment *D1* - Stochastic Gradient Descent Fixed Initial LR approach.



(b) Experiment *E1* - Stochastic Weight Averaging LR finder approach

Figure 6.33: Performance per Age for the Top 3 Best Performers

### 6.6.5 Running Times and Convergence

The runtime is associated to convergence due to the implementation of early stopping for each experiment. Once the loss function ceased to improve with a patience of 10 epochs, the training was stopped. Each optimiser was automatically logged to Neptune and further evaluated for both experiment *D1* and *E1*. The hardware used has a central processing unit (CPU) processor of 2.8GHz (Quad-Core Intel Core i7), memory of 16GB 1600 Mhz DDR3 and an Intel Iris Pro 1536 graphics card.

The SWA optimiser was able to converge the fastest compared to the rest of the algorithms in experiment *D1* with a mean value of approximately 7 minutes. Similar performance occurred in experiment *E1* for which its mean running time was inline with that of SGD with an approximate value of 13 minutes. Despite achieving a low runtime average in *E1*, the SGD algorithm performed the slowest of all algorithms executed for experiment *D1*. This indicates that SGD struggles with a fixed initial LR of `1e−4`.

It is clear that in *D1* the run time average for the different algorithms varied greatly from each other particularly when compared to its *E1* counterpart, which has significantly less dispersion between the mean runtime of the algorithms as shown in Figure 6.34. This suggests that *E1* has a more controlled runtime out of the two experiments. Moreover, with the exception of ADAM, the rest of the algorithms were also found to perform faster in *E1*. These outcomes were due to the automatic LR finder [233], which was made available in experiment *E1*.

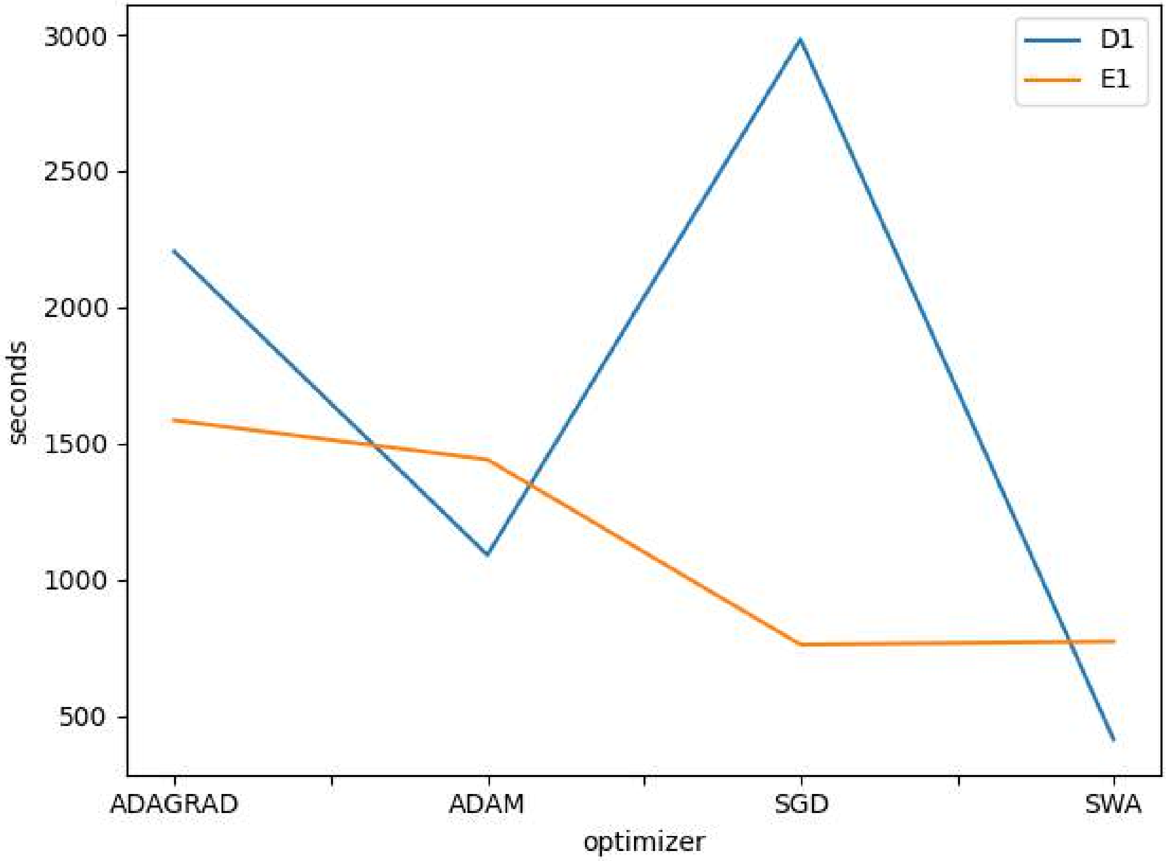


Figure 6.34: Average Runtime per Optimisation Algorithms for Experiments *D1* and *E1*

## 6.7 Summary of Model Results

The summary of the models that were created throughout the research are listed in Table 6.18 with the respective dissemination venues. It can be noticed that the models are designed for specifically the underage age and early adulthood group. Each model is linked to a specific dataset. The specific dataset is an evolution of the data. Unlike the early approach, the size of the dataset has been increasing. This is due to the refinement of the dataset. Whereas the early stage lacked such refinement. The neural networks (NN) used were VGG16, ResNet50 and a 4-Layered NN. Image quality, i.e., blur, noise, exposure and resolution was not measured but image size was recorded. Nevertheless, for the early approach it was not logged. The pre-processing algorithms used were the Azure landmark detector, `dlib` and the DCA approach. The type of algorithms used to solve to age estimation problem were either Classification, Regression or Hybrid. Finally, it was observed that the MAE has been improving from the early approach to the latest facial embedding approach.

| **Report** | **University Report** | **Paper 1 [8]** | **Paper 2 [10]** | **Paper 3 [11]** |
|---|---|---|---|---|
| **Model** | Early Approach | DS13K | DeepUAge | Vec2UAge |
| **Age Range** | 1-25 | 0-25 | 1-20 | 1-18 |
| **Dataset** | VisAGe Beta | DS13K | DeepUAge | Vec2UAge |
| **Dataset Description** | Early Implementation of VisAGe | Flickr + UTKFace + Dataset Generator | VisAGe+ Dataset Generator | VisAGe+ SelfieFV+ Aug. |
| **DS Size** | 17,886 | 12,792 | 17K | 90K |
| **NN** | VGG16 | VGG16 | ResNet50 | 4-L NN |
| **Image Size** | Unknown | 224 x 224 | 224 x 224 | 224 x 224 |
| **Pre-Pro.** | Azure | `Dlib` | DCA | `Dlib` |
| **Algorithm** | C | H | C | R |
| **MAE (Test)** | 6.86 | 3.96 | 2.73 | 2.36 |

Table 6.18: Summary of Created Models. For the Algorithm field, C is Classification, R is Regression and H is Hybrid.

CHAPTER

SEVEN

# CONCLUSIONS AND FUTURE WORK

## 7.1 Summary

Cybercrime Investigations are delayed due to the evergrowing digital forensic backlog. Our main goal is to assist the backlog with DL techniques focused on facial images that can be present in criminal cases involving CSEM. Due to the sensitivity of these cases, using data-driven methods may raise questions in court and if the algorithm does not perform to the satisfaction required, the methods could never be considered again by the Judge.

To address this problem, four research questions were presented in Section 1.3. In this Section, the methods used to address each research question are analysed.

RQ1: “How can the design and implementation of an age-prediction-based DL model aid the digital forensics backlog”

The following hypothesis was presented: H1) DL predicts incriminating content related to child abuse material at a faster rate than human investigators.

Digital Forensic tools require high accuracy. The state-of-the-art methods for facial age estimation were evaluated in Section 4.1 and a baseline analysis of performance evaluation was established. The best methods per year oscillated between MAE values of 2.56 and 6.77. Low performance in age estimation was noticed and a gap in the facial ageing datasets was found. The lack of underage subjects yielded low performance in age estimation. The algorithms developed would be able to surpass the state of the art for underage subjects but a dataset was needed not only to create underage models but also to disseminate an age balanced dataset that could be used for testing and validation. An automated solution was also needed to collect and label single-faced frontal images. The reason the faces were frontal was to divide the problem into a smaller one. Current approaches attempt to generalise the models to different poses, occlusion and low quality tolerance, large age ranges, etc. But this makes the problem highly complex and challenges the performance. Therefore, a smaller problem was addressed. Several approaches, architectures and hyper-parameters were tested. A testing framework was built and the experiments were logged. A regression model based on facial embeddings (outlined in Section 4.5.3) that reached a MAE of 2.36 was achieved, which outperforms other age prediction models.

Additionally, a tool for DF was proposed based on facial embedding extraction. The conventional hash methods to detect CSEM can be improved with the facial embedding approach to recognise faces. This would speed the process of facial recognition and image grouping by subject, in large collections of photographs.

Finally, over 350 experiments were executed with the Vec2UAge approach (outlined in Section 4.5.3). These experiments were logged and each model was produced in an average of 18 minutes. The average runtime per optimisation algorithm was evaluated in Section 6.6.5 and the SWA optimiser was found to be the most optimal.

RQ2: “How can DL be used to aid digital forensic investigators and lessen their exposure to sensitive data?”.

For this research question, the following hypothesis was presented: H1) Many types of ML techniques can discover evidence that is missed by human experts such as pixel patterns or low quality images, i.e., enhanced machine identified features versus those perceptible by human analysts.

Based on the no free lunch theorem, i.e., there is no single model that works best for every problem, our focus was to find a model that suits best for high quality underage age estimation. The approach may fail in other situations, such as adult age ranges or low-quality images. In order to attempt to address the low quality problem, there are several approaches: one of the approaches would be to create a model that is trained with low quality images and employ techniques such as super-resolution processing, de-blurring, etc. Therefore, analysis of low-quality images is addressed in future work. Nevertheless, the improvement of facial age estimation algorithms with the addition of safe technologies, would not only allow the investigator to verify CSEM in a safe manner but it would reduce the exposure and focus on cases were the machine has poor confidence of an image being legal or illegal.

RQ3: “What are the influencing factors that can improve the performance of facial age prediction models”.

Two hypothesis were formulated:

H1) Factors such as emotion, skin tone, facial hair, occlusion, etc., can impact age estimation performance. H2) Gender has a linear correlation with age estimation performance.

To answer this question, several online and offline models were analysed, facial age estimation and attribute prediction was studied. Although ethnicity was a variable that was not analysed in this study, there were several report of studies claiming that algorithms were biased. This was checked in our research. Unbalanced and bias datasets were omnipresent. In our studies several facial attributes were analysed such as hair, makeup, occlusion, quality, skin tone, etc. The skin tone was automatically computed with an FCE algorithm which was explained in Section 3.2.9. It was also determined that there was an influence of gender in age estimation. This was proven in Section 6.1.3. Strong and mild linear correlations were found between the error difference and [exposure, noise, facial hair, gender, anger, glasses, etc].

RQ4: "What is the impact of integrating ML including DL techniques with digital forensic processing with regards to forensic soundness, court admissibility, and case throughput capabilities?".

For this research question, two hypothesis were presented: H1) DL can perform rapid triage to accurately detect illicit multimedia while maintaining forensic soundness. And, H2) ML and DL techniques are not sufficiently accurate for a DF use case.

Many facial age estimation methods are built on datasets with estimated or guessed age label, low sample counts, highly noisy data, unbalanced/biased data, or age bins, e.g., binary groups that separate underage subjects from adults, and/or small groups such as child, teen, adult, etc. Usually, the usage of groups hinder the filtering of illegal content due to its non-dynamic structure. Some of the aforementioned methods affect the reliability of the ML assistance and hence decrease the credibility of the approach. Furthermore, any introduction of reasonable doubt may dismiss a case. As mentioned previously in Section 2.1.2.4, to attempt to alleviate this issue, the Daubert standard is suggested. It was introduced in 1993 and has been used by most state courts in the USA as a rule of evidence to assess the reliability of scientific evidence through the following factors: (1) the method can be and has been tested, (2) subject to peer review, (3) error rates are acceptable, (4) general acceptance in the scientific community. In regards to these factors, Nutter states that "ML easily satisfies three of the four Daubert factors without extensive discussion". However, it should be ensured that the AI in question is not biased in any specific way, especially regarding race/ethnicity and gender.

Intelligent automation is needed to expedite digital investigations that are hampered by lack of resources, such as time and skilled expertise. Moreover, Sanchez et al. through a survey, noted that digital forensic practitioners demand automated tools to detect CSEM, age estimation and skin tone detection. Nevertheless, due to the nature of courtroom practice, and the necessity of expert testimony, it is neither intended nor anticipated that these AI techniques will fully replace trained investigators. Rather, this type of investigative aid has the potential to greatly expedite digital forensic analysts in their work, and potentially lower the psychological load of dealing with CSEM on an ongoing basis

## 7.2 Conclusions

Facial age estimation is still a challenging topic due to external factors such as the environment, habits, ethnicity, diet, etc. Underage age estimation for DF is continuously being studied and the performance has been improving, entitling digital forensic practitioners to use tools and techniques that include Computational Intelligence to detect and analyse evidence; particularly DL. Current models usually attempt to tackle several challenging factors that affect the age estimation performance such as facial occlusion, non-frontal faces, brightness, contrast, quality, etc. In our approach, a simpler challenge is addressed and a better performance is achieved. A distribution for the evaluation model metric is proposed. This distribution is better than having a single value because it allows researchers to chose a model with more confidence. The exploration of optimal LR was key in the influence of high performing underage age estimation models. Another key factor that aided the performance was the quality of the dataset. The creation of the VisAGe dataset enabled the bench-marking of algorithms, analysis of influencing factors on underage facial age estimation, creation of models for the underage group, and the dissemination of the largest human verified underage age estimation dataset. Facial pre-processing techniques are paramount in obtaining better performing models. The DCA approach was developed and a robust technique that produced better results in comparison to other facial pre-processing technologies was presented. The data augmentation techniques have been proved in the past to increase performance but the correct transformation must be chosen; specifically with facial embeddings. The calculation of the facial vectors enabled the use of simpler neural networks. And the experiments were managed swiftly without the use of a GPU. Collaborative tools to record and manage all the experiments, while tracking and visualising metrics such as loss and accuracy, are paramount for researchers not only in DF but in other areas.

The datasets and models created were of predominately white children. It is paramount of being aware of the issues associated with using biased datasets and the harm they can cause as specified in Section 3.4.

## 7.3 Future Work

Our ambition is to investigate how to aid digital forensic cases with automated ML and DL-based techniques. A major problem of the use of ML algorithms is that they operate as black boxes, making it difficult to know how the predictions were achieved. As per Deeks, judges are dealing with an increase of confrontations with ML algorithms in criminal, administrative and civil cases due to the black-box nature of the algorithmic outcomes [57]. The main issue is that the best performing models are the so-called "black boxes" and are not able to explain why certain decision was made [109]. For this reason, the use of Explainable AI (XAI) (which is an AI where the results of a solution can be understood by humans) is imperative and encouraged as supplementary research for law enforcement and courtroom practice.

As future work, our objective is to expand this study further through comparative analysis (with several performance evaluation metrics) of further cloud-based and offline age estimation services. Moreover, the weakness of the current tools is presented where the supplied photograph is not particularly clear. Because of the angle at which it was taken and/or poor quality lighting. These are standard problems in all forms of facial recognition that we wish to address in the future.

We have identified a need for bigger and more diverse datasets for child face recognition to improve our models. Predicting race/ethnicity from images, and the impact of race/ethnicity on the accuracy of predictors using facial images are both topics that have currently received a lot of controversy. These topics are a limitation of our research but of interest for future work.

Once we have collected a dataset with the relevant tags with a considerable size, we would re-train a model specifically for underage images that could help enhance not only age prediction services but also other tools that require identification of child exploitation material.

As future work, there is room to improve the size of VisAGe, expand to more age ranges, emphasise the classification of images by ethnicity, and create specific models that would address certain issues such as low-image quality, illumination and occlusion. The Europol regularly releases unique objects that are present in a crime scene. Technology such as super-resolution could be used to improve the quality of the images and produce better results. This can also be applied to facial images obtained from CCTV cameras to reconstruct an image with better quality and obtain relevant information from the face such as gender, age, skin-tone, ethnicity, etc. It may be worth exploring multi-view approaches to normalise images as a pre-processing step.

In regards to the influencing factors of facial age estimation on underage subjects, further investigation can be conducted on the identification and segregation of negative influencing factors. Exploring the effects of isolating negative influencing factors and the inclusion of only positive influencing factors may have an impact on the accuracy of underage facial age estimation. The coefficient values obtained were based on Pearson's linear approach, it must be considered that a potential strong non-linear correlation may exist between the error difference and another variable. As a result, future work is to explore with nonlinear correlations.

The main goal of this dissertation is to aid law enforcement in the detection and investigation of CSEM. From a victim identification standpoint, we would like to analyse other components that are present in a digital forensic CSEM crime scene including garments, visual geo-location clues, object detection, etc. We also plan to make the VisAGe dataset available and encourage others to contribute, e.g., to improve the demographic balance regarding age/gender/ethnicity that is currently lacking. Lastly, a safe framework will be created to interact with LEAs and evaluate the accuracy of our approaches with real cases.

Finally, while using a fixed encoder (FaceNet) to produce facial embeddings worked well, fine-tuning the encoder by back-propagating the loss from the age estimation module with a low LR should produce a more optimal representation. Trained with a regression loss, the models outlined in the Vec2UAge approach produce a point estimate of the subject's age. However, the value of such an estimate to a DF end-user would be increased if the model instead produced a well-conditioned distribution (mean and variance), for example by applying the methods described in SWA-Gaussian (SWAG) [167]. Lastly, the use of a hyper-parameter optimisation framework for ML such as Optuna [3] would aid the experiments to find improved performance for underage facial age estimation.

# BIBLIOGRAPHY

[1] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, Present, and Future of Face Recognition: A Review. *Electronics*, 9(8), 2020. ISSN 2079-9292. doi: 10.3390/electronics9081188. URL https://www.mdpi.com/2079-9292/9/8/1188.

[2] J. Adserias-Garriga. *Age Estimation: A Multidisciplinary Approach*. Elsevier Science, 2019. ISBN 9780128144923. URL https://books.google.ie/books?id=bPKRDwAAQBAJ.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631, 2019.

[4] Noora Ahmad Khurshid Al Mutawa. *Integrating Behavioural Analysis within the Digital Forensics Investigation Process*. PhD thesis, University of Central Lancashire, 2018.

[5] Ernie Allen. Child Pornography: Model Legislation & Global Review. *Accessed on 13th August*, 2007.

[6] F. Anda, D. Lillis, N. Le-Khac, and M. Scanlon. Evaluating Automated Facial Age Estimation Techniques for Digital Forensics. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 129–139, 2018.

[7] Felix Anda, Nhien-An Le-Khac, and Mark. Scanlon. Poster: Automated Machine Learning-Based Digital Evidence Classification Techniques. In *16th European Conference on Cyber Warfare and Security (EC-CWS)*, Dublin, Ireland, 2017.

[8] Felix Anda, David Lillis, Aikaterini Kanta, Brett A Becker, Elias Bou-Harb, Nhien-An Le-Khac, and Mark Scanlon. Improving the accuracy of automated facial age estimation to aid CSEM investigations. *Digital Investigation*, 2019.

[9] Felix Anda, Brett A Becker, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. Assessing the influencing factors on the accuracy of underage facial age estimation. In *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, pages 1–8. IEEE, 2020.

[10] Felix Anda, Nhien-An Le-Khac, and Mark Scanlon. DeepUAge: Improving Underage Age Estimation Accuracy to Aid CSEM Investigation. *Forensic Science International: Digital Investigation*, 32:300921, 2020. ISSN 2666-2817. doi: https://doi.org/10.1016/j.fsidi.2020.300921. URL `http://www.sciencedirect.com/science/article/pii/S2666281720300160`.

[11] Felix Anda, Edward Dixon, Elias Bou-Harb, Nhien-an Le-Khac, and Mark Scanlon. Vec2uage: Enhancing underage age estimation performance through facial embeddings. *Forensic Science International: Digital Investigation*, 2021.

[12] Félix Anda Basabe. Forensic Data Recovery from Android Smart Watches. Master's thesis, King's College London, 2016.

[13] Raphael Angulu, Jules R Tapamo, and Aderemi O Adewumi. Age estimation via face images: a survey. *EURASIP Journal on Image and Video Processing*, 2018(1):42, 2018.

[14] Grigory Antipov, Moez Baccouche, Sid-Ahmed Berrani, and Jean-Luc Dugelay. Apparent age estimation from face images combining general and children-specialized deep learning models. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 96–104, 2016.

[15] Rigan Ap-Apid. An algorithm for nudity detection. In *5th Philippine Computing Science Congress*, pages 201–205, 2005.

[16] London (United Kingdom); Audit Commission. *Ghost in the machine An analysis of IT fraud and abuse*. Audit Commission, 1998.

[17] AWSChris. Detecting Faces in an Image. `https://github.com/awsdocs/amazon-rekognition-developer-guide/blob/master/doc_source/faces-detect-images.md`, 2020.

[18] John Bandler and Antonia Merzon. *Cybercrime Investigations: A Comprehensive Resource for Everyone*. CRC Press, 2020.

[19] Vivek Bannore. *Introduction to Super-Resolution*, pages 1–8. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-00385-1. doi: 10.1007/978-3-642-00385-1_1. URL `https://doi.org/10.1007/978-3-642-00385-1_1`.

[20] C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483, 1996.

[21] Belhassen Bayar and Matthew C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10. ACM, 2016.

[22] RE Bell. The prosecution of computer crime. *Journal of financial crime*, 9(4):308–325, 2002.

[23] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[24] August Bequai. A guide to cyber-crime investigations, 1998.

[25] Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2): 100205, 2021.

[26] Marcus D Bloice, Peter M Roth, and Andreas Holzinger. Biomedical image augmentation using Augmentor. *Bioinformatics*, 35(21):4522–4524, 04 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz259.

[27] Richard Boddington. A case study of the challenges of cyber forensics analysis of digital evidence in a child pornography trial. In *Proceedings of the Conference on Digital Forensics, Security and Law*, page 155. Association of Digital Forensics, Security and Law, 2012.

[28] Timothy Bollé, Eoghan Casey, and Maëlig Jacquet. The role of evaluations in reaching decisions using automated systems Supporting forensic analysis kk. *Forensic Science International*, 34:301016, 2020.

[29] Ranjit Bose, Xin Robert Luo, and Yuan Liu. The roles of security and trust: Comparing cloud computing and banking. *Procedia-Social and Behavioral Sciences*, 73: 30–34, 2013.

[30] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[31] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[32] Kevin W Bowyer. Face recognition technology: security versus privacy. *IEEE Technology and society magazine*, 23(1):9–19, 2004.

[33] Susan W Brenner, Brian Carrier, and Jef Henninger. The Trojan horse defense in cybercrime cases. *Santa Clara Computer & High Tech. LJ*, 21:1, 2004.

[34] Vicki Bruce, A Mike Burton, Elias Hanna, Pat Healey, Oli Mason, Anne Coombes, Rick Fright, and Alf Linney. Sex discrimination: how do we tell the difference between male and female faces? *perception*, 22(2):131–152, 1993.

[35] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[36] B Buyuktas, Cigdem Eroglu Erdem, and AT Erdem. Curriculum Learning for Face Recognition. In *European Signal Processing Conference (EUSIPCO)*, 2020.

[37] E. Buza, A. Akagic, and S. Omanovic. Skin detection based on image color segmentation with histogram and K-means clustering. In *2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pages 1181–1186, Nov 2017.

[38] Brian Carrier. Autopsy, 2015.

[39] John Carter and Mark Nixon. An integrated biometric database. In *IEE Colloquium on Electronic Images and Image Processing in Security and Forensic Science*, pages 8–1. IET, 1990.

[40] Eoghan Casey, Monique Ferraro, and Lam Nguyen. Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence. *Journal of forensic sciences*, 54(6):1353–1364, 2009.

[41] Modesto Castrillón-Santana, José Javier Lorenzo Navarro, and Cristina Freire Obregón. Boys2Men, an age estimation dataset with applications to detect enfants in pornography content, 2016.

[42] Oya Çeliktutan, Sezer Ulukaya, and Bülent Sankur. A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, 2013(1): 13, 2013.

[43] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *CVPR 2011*, pages 585–592. IEEE, 2011.

[44] Wei-Lun Chao, Jun-Zuo Liu, and Jian-Jiun Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628 – 641, 2013. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2012.09.011.

[45] Cunjian Chen, Antitza Dantcheva, and Arun Ross. Impact of facial cosmetics on automatic gender and age estimation algorithms. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 182–190. IEEE, 2014.

[46] Jiansheng Chen, Xiangui Kang, Ye Liu, and Z Jane Wang. Median filtering forensics based on convolutional neural networks. *IEEE Signal Processing Letters*, 22(11):1849–1853, 2015.

[47] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013.

[48] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[49] T.K. Clancy. *Cyber Crime and Digital Evidence: Materials and Cases*. LexisNexis, 2011. ISBN 9780327175865. URL `https://books.google.ie/books?id=xbM5kONbnloC`.

[50] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6): 681–685, 2001.

[51] M. d. Polastro and P. M. da Silva Eleuterio. NuDetective: A Forensic Tool to Help Combat Child Pornography through Automatic Nudity Detection. In *2010 Workshops on Database and Expert Systems Applications*, pages 349–353, 2010. doi: 10.1109/DEXA.2010.74.

[52] Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu. Eigentransfer: A unified framework for transfer learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 193–200, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553399. URL `http://doi.acm.org/10.1145/1553374.1553399`.

[53] Kirsten A Dalrymple, Jesse Gomez, and Brad Duchaine. The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. *PloS one*, 8(11):e79131, 2013.

[54] A. Dantcheva, C. Chen, and A. Ross. Can facial cosmetics affect the matching accuracy of face recognition systems? In *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 391–398, Sep. 2012. doi: 10.1109/BTAS.2012.6374605.

[55] Antitza Dantcheva, Carmelo Velardo, Angela D'Angelo, and Jean-Luc Dugelay. Bag of soft biometrics for person identification. *Multimedia Tools and Applications*, 51(2):739–777, Jan 2011. ISSN 1573-7721. doi: 10.1007/s11042-010-0635-7. URL `https://doi.org/10.1007/s11042-010-0635-7`.

[56] Debayan Deb, Neeta Nain, and Anil K Jain. Longitudinal study of child face recognition. In *2018 International Conference on Biometrics (ICB)*, pages 225–232. IEEE, 2018.

[57] Ashley Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7):1829–1850, 2019.

[58] Afshin Dehghan, Enrique G Ortiz, Guang Shu, and Syed Zain Masood. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*, 2017.

[59] Hedwige Dehon and Serge Brédart. An 'other-race'effect in age estimation from faces. *Perception*, 30(9):1107–1113, 2001.

[60] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[61] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[62] Thomas G Dietterich. Machine-learning research. *AI magazine*, 18(4):97–97, 1997.

[63] Edward Dixon. Selfie-FV: Face vectors with age ground-truth. https://github.com/EdwardDixon/selfie-fv, 2020.

[64] Yuan Dong, Yinan Liu, and Shiguo Lian. Automatic age estimation based on deep learning algorithm. *Neurocomputing*, 187:4–10, 2016.

[65] Paul Steven Dowland. *User authentication and supervision in networked systems*. PhD thesis, University of Plymouth, 2004.

[66] Xiaoyu Du and Mark Scanlon. Methodology for the Automated Metadata-Based Classification of Incriminating Digital Forensic Artefacts. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, ARES '19, pages 43:1–43:8, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-7164-3. doi: 10.1145/3339252.3340517. URL http://doi.acm.org/10.1145/3339252.3340517.

[67] Xiaoyu Du, Nhien-An Le-Khac, and Mark Scanlon. Evaluation of Digital Forensic Process Models with Respect to Digital Forensics as a Service. In *Proceedings of the 16th European Conference on Cyber Warfare and Security (ECCWS 2017)*, pages 573–581. ACPI, 2017.

[68] Xiaoyu Du, Chris Hargreaves, John Sheppard, Felix Anda, Asanka Sayakkara, Nhien-An Le-Khac, and Mark Scanlon. SoK: Exploring the State of the Art and the Future Potential of Artificial Intelligence in Digital Forensic Investigation. In *Proceedings of the 15th International Conference on Availability, Reliability and Security*, ARES '20, pages 1–10, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450388337. doi: 10.1145/3407023.3407068. URL https://doi.org/10.1145/3407023.3407068.

[69] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[70] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

[71] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[72] Europol. Criminal networks involved in the trafficking and exploitation of underage victims in the European Union, 2018. URL https://www.europol.europa.eu/publications-documents/criminal-networks-involved-in-trafficking-and-exploitation-of-underage-victims-in-eu.

[73] European Cybercrime Center (EC3) Europol. Commercial Sexual Exploitation of Children Online, 2013. URL `https://www.europol.europa.eu/sites/default/files/documents/efc_strategic_assessment_2014.pdf`.

[74] Hany Farid. Reining in Online Abuses. *Technology & Innovation*, 19(3):593–599, 2018.

[75] Jason Farina, Mark Scanlon, Nhien-An Le-Khac, and M-Tahar Kechadi. Overview of the Forensic Investigation of Cloud Services. In *10th International Conference on Availability, Reliability and Security (ARES 2015)*, pages 556–565, Toulouse, France, 08 2015. IEEE. doi: 10.1109/ARES.2015.81.

[76] Mohd Awais Farooque and Jayant S Rohankar. Survey on various noises and techniques for denoising the color image. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2(11):217–221, 2013.

[77] Eilidh Ferguson and Caroline Wilkinson. Juvenile age estimation from facial images. *Science & Justice*, 57(1):58 – 62, 2017. ISSN 1355-0306. doi: https://doi.org/10.1016/j.scijus.2016.08.005. URL `http://www.sciencedirect.com/science/article/pii/S1355030616300739`.

[78] Carles Fernández, Ivan Huerta, and Andrea Prati. A comparative evaluation of regression learning algorithms for facial age estimation. In *International Workshop on Face and Facial Expression Recognition from Real World Videos*, pages 133–144. Springer, 2014.

[79] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. Nist special databse 32-multiple encounter dataset ii (meds-ii). *NIST Interagency/Internal Report (NISTIR)-7807*, 2011.

[80] BT Fraser. Computer Crime Research Resources. *School of Library and Information Studies, Florida State University,¡ http://mailer. fsu. edu/˜ btf1553/ccrr/search*, 1, 1996.

[81] David Freire-Obregon, Fabio Narducci, Silvio Barra, and Modesto Castrillon-Santana. Deep learning for source camera identification on mobile devices. *Pattern Recognition Letters*, 2018. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2018.01.005. URL `http://www.sciencedirect.com/science/article/pii/S0167865518300059`.

[82] Y. Fu, G. Guo, and T. S. Huang. Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, Nov 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2010.36.

[83] SM Furnell. The problem of categorising cybercrime and cybercriminals. In *2nd Australian information warfare and security conference*, 2001.

[84] Matt Fussell. Facial Proportions - How to Draw a Face. `https://thevirtualinstructor.com/facialproportions.html`, 2019. [Online; accessed 04-10-2019].

[85] Tzvi Ganel. Smiling makes you look older. *Psychonomic bulletin & review*, 22(6): 1671–1677, 2015.

[86] Clare Garvie and Jonathan Frankle. Facial-recognition software might have a racial bias problem. *The Atlantic*, 7, 2016.

[87] X. Geng, C. Yin, and Z. Zhou. Facial Age Estimation by Learning from Label Distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (10):2401–2412, 2013.

[88] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 307–316, 2006.

[89] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. SEXNET: A Neural Network Identifies Sex from Human Faces. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems*, NIPS'90, page 572, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc. ISBN 1558601848.

[90] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[91] Sean E Goodison, Robert C Davis, and Brian A Jackson. Digital evidence and the us criminal justice system. *Identifying Technology and Other Needs to More Effectively Acquire and Utilize Digital Evidence. Priority Criminal Justice Needs Initiative. Rand Corporation*, 2015.

[92] Google. Google Trend of the ”Deep Learning” search term for the past 5 years, 2020. URL `https://trends.google.com/trends/explore?date=today5-y&q=deeplearning,machinelearning,artificialintelligence`. [Online; accessed 30-11-2020].

[93] Sarah Gordon and Richard Ford. On the definition and classification of cybercrime. *Journal in Computer Virology*, 2(1):13–20, 2006.

[94] John C Gower. Generalized Procrustes Analysis. *Psychometrika*, 40(1):33–51, 1975.

[95] Cinthya Grajeda, Frank Breitinger, and Ibrahim Baggili. Availability of datasets for digital forensics–And what is missing. *Digital Investigation*, 22:S94–S105, 2017.

[96] Petra Grd and Miroslav Bača. Creating a face database for age estimation and classification. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on*, pages 1371–1374. IEEE, 2016.

[97] G. Guo and G. Mu. Simultaneous dimensionality reduction and human age estimation via kernel partial least squares regression. In *CVPR 2011*, pages 657–664, 2011. doi: 10.1109/CVPR.2011.5995404.

[98] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang. A Study on Automatic Age Estimation Using a Large Database. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1986–1991, Sep. 2009. doi: 10.1109/ICCV.2009.5459438.

[99] Guodong Guo and Guowang Mu. Human age estimation: What is the influence across race and gender? In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 71–78. IEEE, 2010.

[100] Guodong Guo and Guowang Mu. Joint estimation of age, gender and ethnicity: Cca vs. pls. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–6. IEEE, 2013.

[101] S.L. Hamilton. *Forensic Artist: Solving the Case with a Face*. Crime Scene Investigation. ABDO Publishing Company, 2010. ISBN 9781617842726. URL `https://books.google.ie/books?id=J5p7AgAAQBAJ`.

[102] H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic Estimation from Face Images: Human vs. Machine Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1148–1161, June 2015. doi: 10.1109/TPAMI.2014.2362759.

[103] Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 87:27, 2014.

[104] Hu Han, Charles Otto, and Anil K Jain. Age estimation from face images: Human vs. machine performance. In *Biometrics (ICB), 2013 International Conference on*, pages 1–8. IEEE, 2013.

[105] Ronald Henss. Perceiving age and attractiveness in facial photographs. *Journal of Applied Social Psychology*, 21(11):933–946, 1991.

[106] Allan Heydon and Marc Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–229, 1999.

[107] Leo Hidalgo. Update on Creative Commons licenses and "In Memoriam" accounts., 2019. URL `https://blog.flickr.net/2019/03/08/update-on-creative-commons-licenses-and-in-memoriam-accounts/`. [Online; accessed 30-11-2020].

[108] Ben Hitchcock, Nhien-An Le-Khac, and Mark Scanlon. Tiered forensic methodology model for digital field triage by non-digital evidence specialists. *Digital Investigation*, 16:S75 – S85, 2016. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2016.01.010. URL `http://www.sciencedirect.com/science/article/pii/S1742287616300044`. DFRWS 2016 Europe.

[109] Andreas Holzinger. From machine learning to explainable AI. In *2018 world symposium on digital intelligence for systems and machines (DISA)*, pages 55–66. IEEE, 2018.

[110] Wen-Bing Horng, Cheng-Ping Lee, and Chun-Wen Chen. Classification of age groups based on facial features. *Tamkang Journal of Science and Engineering*, 4(3): 183–191, 2001.

[111] Satoshi Hosoi, Erina Takikawa, and Masato Kawade. Ethnicity estimation with facial images. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 195–200. IEEE, 2004.

[112] Zhenzhen Hu, Yonggang Wen, Jianfeng Wang, Meng Wang, Richang Hong, and Shuicheng Yan. Facial age estimation with age difference. *IEEE Transactions on Image Processing*, 26(7):3087–3097, 2016.

[113] Wenyi Huang, Ingmar Weber, and Sarah Vieweg. Inferring nationalities of twitter users and studying inter-national linking. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 237–242, 2014.

[114] Paul Hunton. A rigorous approach to formalising the technical investigation stages of cybercrime and criminality within a uk law enforcement environment. *Digital Investigation*, 7(3):105 – 113, 2011. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2011.01.002. URL `http://www.sciencedirect.com/science/article/pii/S174228761100003X`.

[115] Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J Walls, Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, Brian Lynn, and Janis Wolak. Measurement and analysis of child pornography trafficking on P2P networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 631–642. ACM, 2013.

[116] ECPAT International. Trends in online child sexual abuse material. *ECPAT International*, 2018. URL `https://www.ecpat.org/wp-content/uploads/2018/07/ECPAT-International-Report-Trends-in-Online-Child-Sexual-Abuse-Material-2018.pdf`.

[117] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[118] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A.G. Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In *34th Conference on Uncertainty in Artificial Intelligence*, volume 2, pages 876–885, 2018.

[119] Bastian Jaeger, Willem WA Sleegers, and Anthony M Evans. Automated classification of demographics from face images: A tutorial and validation. *Social and Personality Psychology Compass*, 14(3):e12520, 2020.

[120] Anil K Jain, Sarat C Dass, and Karthik Nandakumar. Can soft biometric traits assist user recognition? In *Biometric technology for human identification*, volume 5404, pages 561–572. International Society for Optics and Photonics, 2004.

[121] Joshua I James and Pavel Gladyshev. Challenges with Automation in Digital Forensic Investigations. *arXiv preprint arXiv:1303.4498*, 2013.

[122] Joshua I James and Pavel Gladyshev. Automated inference of past action instances in digital investigations. *International Journal of Information Security*, 14 (3):249–261, 2015.

[123] Stuart H James and Jon J Nordby. *Forensic science: an introduction to scientific and investigative techniques*. CRC press, 2002.

[124] H Marshall Jarrett, Michael W Bailie, E Hagen, and E Etringham. Prosecuting computer crimes. *Office of Legal Education Executive Office for United States Attorneys Accessible from US Justice Department. Retrieved October*, 27:2015, 2010.

[125] Charles F. Jekel and Raphael T. Haftka. Classifying Online Dating Profiles on Tinder using FaceNet Facial Embeddings. *CoRR*, abs/1803.04347, 2018. URL http://arxiv.org/abs/1803.04347.

[126] Jennifer Lee Jenkins, Melissa L. McCarthy, Lauren M. Sauer, Gary B. Green, Stephanie Stuart, Tamara L. Thomas, and Edbert B. Hsu. Mass-Casualty Triage: Time for an Evidence-Based Approach. *Prehospital and Disaster Medicine*, 23(1): 3–8, 2008. doi: 10.1017/S1049023X00005471.

[127] Philip Jenkins et al. *Beyond tolerance: Child pornography on the Internet*. NYU Press, 2001.

[128] Yvonne Jewkes and Carol Andrews. Policing the filth: The problems of investigating online child pornography in England and Wales. *Policing and Society*, 15 (1):42–62, 2005. doi: 10.1080/1043946042000338922. URL https://doi.org/10.1080/1043946042000338922.

[129] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[130] Brian Jones, Syd Pleno, and Michael Wilkinson. The use of random sampling in investigations involving child abuse material. *Digital Investigation*, 9:S99 – S107, 2012. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2012.05.011. URL http://www.sciencedirect.com/science/article/pii/S1742287612000369. The Proceedings of the Twelfth Annual DFRWS Conference.

[131] Kimmo Karkkainen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558, January 2021.

[132] R Katoch and S Rajagopalan. Warfare injuries: History, triage, transport and field hospital setup in the armed forces. *Medical Journal Armed Forces India*, 66(4): 304–308, 2010.

[133] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[134] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016.

[135] Orin S Kerr. Vagueness Challenges to the Computer Fraud and Abuse Act. *Minn. L. Rev.*, 94:1561, 2009.

[136] UA Khan, MI Cheema, and NM Sheikh. Adaptive video encoding based on skin tone region detection. In *IEEE Students Conference, ISCON'02. Proceedings.*, volume 1, pages 129–134. IEEE, 2002.

[137] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[138] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[139] Juliane A Kloess, Jessica Woodhams, Helen Whittle, Tim Grant, and Catherine E Hamilton-Giachritsis. The challenges of identifying and classifying child sexual abuse material. *Sexual Abuse*, page 1079063217724768, 2017.

[140] Jochem Kossen. Showoff - Webbased photo management. `https://github.com/jkossen/showoff`, 2020.

[141] Rafał Kozik, Michał Choraś, and Witold Hołubowicz. Evolutionary-based packets classification for anomaly detection in web layer. *Security and Communication Networks*, 9(15):2901–2910, 2016.

[142] Meredith Krause. Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology*, 24(1): 22–29, 2009.

[143] Tony Krone. A Typology of Online Child Pornography Offending. *Trends and Issues in Crime and Criminal Justice*, 279, 01 2004.

[144] Warren G Kruse II and Jay G Heiser. *Computer forensics: incident response essentials*. Pearson Education, 2001.

[145] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23:1189–1197, 2010.

[146] Young H Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1 – 21, 1999. ISSN 1077-3142. doi: https://doi.org/10.1006/cviu.1997.0549. URL `http://www.sciencedirect.com/science/article/pii/S107731429790549X`.

[147] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[148] Andreas Lanitis, Chrisina Draganova, and Chris Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

[149] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[150] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR 2011*, pages 1721–1728. IEEE, 2011.

[151] Fwa Hua Leong. Deep learning of facial embeddings and facial landmark points for the detection of academic emotions. In *Proceedings of the 5th International Conference on Information and Education Innovations*, pages 111–116, 2020.

[152] Anna Leppänen and Terhi Kankaanranta. Cybercrime investigation in Finland. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, 18(2):157–175, 2017.

[153] Gil Levi and Tal Hassner. Age and Gender Classification Using Convolutional Neural Networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*, June 2015.

[154] Changsheng Li, Qingshan Liu, Weishan Dong, Xiaobin Zhu, Jing Liu, and Hanqing Lu. Human age estimation based on locality and ordinal information. *IEEE transactions on cybernetics*, 45(11):2522–2534, 2014.

[155] Haibin Liao, Yuchen Yan, Wenhua Dai, and Ping Fan. Age estimation of face images based on CNN and divide-and-rule strategy. *Mathematical Problems in Engineering*, 2018, 2018.

[156] David Lillis, Brett Becker, Tadhg O'Sullivan, and Mark Scanlon. Current Challenges and Future Research Areas for Digital Forensic Investigation. In *The 11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016)*, pages 9–20, Daytona Beach, FL, USA, 05 2016. ADFSL.

[157] David Lillis, Frank Breitinger, and Mark Scanlon. Hierarchical Bloom filter trees for approximate matching. *The Journal of Digital Forensics, Security and Law: JDFSL*, 13(1):81–96, 2018.

[158] Hao Liu, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Group-aware deep feature learning for facial age estimation. *Pattern Recognition*, 66:82 – 94, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2016.10.026.

[159] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. AgeNet: Deeply Learned Regressor and Classifier for Robust Apparent Age Estimation. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 258–266, Dec 2015. doi: 10.1109/ICCVW.2015.42.

[160] Xinhua Liu, Yao Zou, Hailan Kuang, and Xiaolin Ma. Face Image Age Estimation Based on Data Augmentation and Lightweight Convolutional Neural Network. *Symmetry*, 12(1):146, 2020.

[161] Y. Liu, Yongzheng Lin, and Y. Chen. Ensemble classification based on ica for face recognition. *2008 Congress on Image and Signal Processing*, 3:144–148, 2008.

[162] Steve Lohr. Facial recognition is accurate, if you're a white guy. *New York Times*, 9, 2018.

[163] Andrew Loomis. *Figure Drawing for All It's Worth*. Editoria Bibliomundi Serviçis Digitais LTDA, 2017.

[164] Xiaoguang Lu, Anil K Jain, et al. Ethnicity identification from face images. In *Proceedings of SPIE*, volume 5404, pages 114–123. Citeseer, 2004.

[165] Khoa Luu, Keshav Seshadri, Marios Savvides, Tien D Bui, and Ching Y Suen. Contourlet appearance model for facial age estimation. In *2011 international joint conference on biometrics (IJCB)*, pages 1–8. IEEE, 2011.

[166] Jiang-Jing Lv, Xiao-Hu Shao, Jia-Shui Huang, Xiang-Dong Zhou, and Xi Zhou. Data augmentation for face recognition. *Neurocomputing*, 230:184 – 196, 2017. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2016.12.025.

[167] Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13153–13164. Curran Associates, Inc., 2019.

[168] Corey Manders, Farzam Farbiz, and Chong Jyh Herng. The effect of linearization of range in skin detection. In *2007 6th International Conference on Information, Communications & Signal Processing*, pages 1–5. IEEE, 2007.

[169] Gabriela E Martínez, Patricia Melin, Olivia D Mendoza, and Oscar Castillo. Face recognition with a Sobel edge detector and the Choquet integral as integration method in a modular neural networks. In *Design of intelligent systems based on fuzzy logic, neural networks and nature-inspired optimization*, pages 59–70. Springer, 2015.

[170] Fabio Marturana and Simone Tacconi. A machine learning-based triage methodology for automated categorization of digital media. *Digital Investigation*, 10(2):193 – 204, 2013. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2013.01.001. URL `http://www.sciencedirect.com/science/article/pii/S1742287613000029`. Triage in Digital Forensics.

[171] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Gozde Sahin, and Gérard Medioni. Face-specific data augmentation for unconstrained face recognition. *International Journal of Computer Vision*, 127(6-7):642–667, 2019.

[172] Mike McGuire and Samantha Dowling. Cyber-crime: A review of the evidence Research Report 75, Chapter 2: Cyber-enabled crimes-fraud and theft. *Home Office*, pages 1–27, 2013.

[173] Rodney McKemmish. When is digital evidence forensically sound? In *IFIP international conference on digital forensics*, pages 3–15. Springer, 2008.

[174] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. Diversity in faces. *arXiv preprint arXiv:1901.10436*, 2019.

[175] Microsoft. Azure Cognitive Services Face API Documentation, 2020. URL `https://docs.microsoft.com/en-us/azure/cognitive-services/face/`. [Online; accessed 1-12-2020].

[176] Stephen Milborrow and Fred Nicolls. Locating facial features with an extended active shape model. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 504–513, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-88693-8.

[177] Garda Missing Persons Bureau. Guidance on the Recording Investigation and Management of Missing Persons, 2020. URL `https://www.garda.ie/en/about-us/publications/policy-documents/guidance-on-the-recording-investigation-and-management-of-missing-persons.pdf`.

[178] Ahmad Saeed Mohammad and Jabir Alshehabi Al-Ani. Towards ethnicity detection using learning based classifiers. In *2017 9th Computer Science and Electronic Engineering (CEEC)*, pages 219–224. IEEE, 2017.

[179] E. Moyse and S. Brédart. An own-age bias in age estimation of faces. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, 62(1):3 – 7, 2012. ISSN 1162-9088. doi: https://doi.org/10.1016/j.erap.2011.12.002.

[180] Evelyne Moyse. Age estimation from faces and voices: a review. *Psychologica Belgica*, 54(3):255–265, 2014.

[181] Martin Mulazzani. New challenges in digital forensics: online storage and anonymous communication. *Diss. Vienna University of Technology*, 2014.

[182] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 2012. ISBN 9780262018029. URL `https://books.google.ie/books?id=NZP6AQAAQBAJ`.

[183] Muhammad Shahroz Nadeem, Virginia NL Franqueira, and Xiaojun Zhai. *Privacy verification of PhotoDNA based on machine learning*. IET, 2019. doi: 10.1049/PBPC028E_ch12.

[184] John Nawara. Machine Learning: Face Recognition Technology Evidence in Criminal Trials. *U. Louisville L. Rev.*, 49:601, 2010.

[185] Vivens Ndatinya, Zhifeng Xiao, Vasudeva Rao Manepalli, Ke Meng, and Yang Xiao. Network forensics analysis using Wireshark. *International Journal of Security and Networks*, 10(2):91–106, 2015.

[186] neptune.ai. Neptune: experiment management and collaboration tool, 2020. URL `https://neptune.ai`.

[187] Adrian Nestor and Michael J. Tarr. Gender Recognition of Human Faces Using Color. *Psychological Science*, 19(12):1242–1246, 2008. doi: 10.1111/j.1467-9280.2008.02232.x. URL `https://doi.org/10.1111/j.1467-9280.2008.02232.x`. PMID: 19121131.

[188] Fudong Nian, Teng Li, Yan Wang, Mingliang Xu, and Jun Wu. Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing*, 210:283 – 293, 2016. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2015.09.135. URL `http://www.sciencedirect.com/science/article/pii/S0925231216305963`. SI:Behavior Analysis In SN.

[189] Mariam Nouh, Jason RC Nurse, Helena Webb, and Michael Goldsmith. Cybercrime Investigators are Users Too! Understanding the Socio-Technical Challenges Faced by Law Enforcement. *arXiv preprint arXiv:1902.06961*, 2019.

[190] Human Rights Now. Report on child pornography in Japan, 2018. URL `http://hrn.or.jp/eng/wp-content/uploads/2018/02/HRN-Child-Pornography-Report-2018.02.07.pdf`.

[191] Patrick W Nutter. Machine Learning Evidence: Admissibility and Weight. *U. Pa. J. Const. L.*, 21:919, 2018.

[192] Office of Juvenile Justice and Delinquency Prevention. When your child is missing: A family survival guide, 2010.

[193] United Nations Office on Drugs and Crime Vienna. Comprehensive Study on Cybercrime, 2013. URL `https://www.unodc.org/documents/organized-crime/UNODC_CCPCJ_EG.4_2013/CYBERCRIME_STUDY_210213.pdf`.

[194] United Nations Office on Drugs and Crime Vienna. Study on the Effects of New Information Technologies on the Abuse and Exploitation of Children, 2015. URL `https://www.unodc.org/documents/Cybercrime/Study_on_the_Effects.pdf`.

[195] Richard E Overill and Jantje A. M. Silomon. Single and Double Power Laws for Cyber-Crimes. *Journal of Information Warfare*, 10(3):29–36, 2011. ISSN 14453312, 14453347. URL `https://www.jstor.org/stable/26486815`.

[196] Richard E. Overill, Jantje A M Silomon, and Keith A. Roscoe. Triage template pipelines in digital forensic investigations. *Digital Investigation*, 10(2):168–174, 9 2013. ISSN 1742-2876. doi: 10.1016/j.diin.2013.03.001.

[197] Gary Palmer et al. A road map for digital forensic research. In *First digital forensic research workshop, utica, new york*, pages 27–30, 2001.

[198] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[199] Divyarajsinh N Parmar and Brijesh B Mehta. Face Recognition Methods & Applications. *International Journal of Computer Technology and Applications*, 4(1):84–86, 2013.

[200] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[201] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998.

[202] John B Pittenger and Robert E Shaw. Perception of relative and absolute age in facial photographs. *Attention, Perception, & Psychophysics*, 18(2):137–143, 1975.

[203] Andrée Pomerleau, Daniel Bolduc, Gérard Malcuit, and Louise Cossette. Pink or blue: Environmental gender stereotypes in the first two years of life. *Sex Roles*, 22(5):359–367, Mar 1990. ISSN 1573-2762. doi: 10.1007/BF00288339. URL `https://doi.org/10.1007/BF00288339`.

[204] Darren Quick and Kim-Kwang Raymond Choo. Impacts of increasing volume of digital forensic data: A survey and future research challenges. *Digital Investigation*, 11(4):273–294, 2014.

[205] KVN Rajesh and KVN Ramesh. Artificial Intelligence–Fact or Fiction. *Computing NaNo*, 2012.

[206] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 145–151, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375820. URL `https://doi-org.ucd.idm.oclc.org/10.1145/3375627.3375820`.

[207] N. Ramanathan and R. Chellappa. Modeling Age Progression in Young Faces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 387–394, June 2006. doi: 10.1109/CVPR.2006.187.

[208] Yuan Rao and Jiangqun Ni. A deep learning approach to detection of splicing and copy-move forgeries in images. In *Information Forensics and Security (WIFS), 2016 IEEE International Workshop on*, pages 1–6. IEEE, 2016.

[209] M Ratnayake, Z Obertová, M Dose, P Gabriel, HM Bröker, M Brauckmann, A Barkus, R Rizgeliene, J Tutkuviene, Stefanie Ritz-Timme, et al. The Juvenile Face as a Suitable Age Indicator in Child Pornography Cases: A Pilot Study on

the Reliability of Automated and Visual Estimation Approaches. *International journal of legal medicine*, 128(5):803–808, 2014.

[210] Daniel Reisfeld and Yehezkel Yeshurun. Preprocessing of Face Images: Detection of Features and Pose Normalization. *Computer Cision and Image Understanding*, 71 (3):413–430, 1998.

[211] Lauren Rhue. Who gets started on kickstarter? demographic variations in fundraising success. In *ICIS*, 2015.

[212] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 341–345. IEEE, 2006.

[213] Karl Ricanek, Shivani Bhardwaj, and Michael Sodomsky. A Review of Face Recognition against Longitudinal Child Faces. In Arslan Brömme, Christoph Busch, Christian Rathgeb, and Andreas Uhl, editors, *Proceedings of the 14th International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 15–26, Bonn, 2015. Gesellschaft für Informatik e.V.

[214] T Rimo and M Walth. McAfee and CSIS: Stopping Cybercrime Can Positively Impact World Economies. *McAfee, June*, 9, 2014.

[215] Iain Robertson-Steel. Evolution of triage systems. *Emergency Medicine Journal*, 23 (2):154–155, 2006.

[216] Pau Rodríguez, Guillem Cucurull, Josep M Gonfaus, F Xavier Roca, and Jordi Gonzàlez. Age and gender recognition in the wild with deep attention. *Pattern Recognition*, 72:563–571, 2017.

[217] Marcus K Rogers, James Goldman, Rick Mislan, Timothy Wedge, and Steve Debrota. Computer forensics field triage process model. In *Proceedings of the conference on Digital Forensics, Security and Law*, page 27. Association of Digital Forensics, Security and Law, 2006.

[218] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4):144–157, 2018.

[219] V. Roussev. Hashing and Data Fingerprinting in Digital Forensics. *IEEE Security Privacy*, 7(2):49–55, 2009. doi: 10.1109/MSP.2009.40.

[220] Stuart J Russell and Peter Norvig. Artificial Intelligence: A Modern Approach, 2012.

[221] Napa Sae-Bae, Xiaoxi Sun, Husrev T Sencar, and Nasir D Memon. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5332–5336. IEEE, 2014.

[222] Richard P Salgado. Fourth amendment search and the power of the hash. *Harv. L. Rev. F.*, 119:38, 2005.

[223] Laura Sanchez, Cinthya Grajeda, Ibrahim Baggili, and Cory Hall. A Practitioner Survey Exploring the Value of Forensic Tools, AI, Filtering, & Safer Presentation for Investigating Child Sexual Abuse Material (CSAM). *Digital Investigation*, 29:S124 – S142, 2019. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2019.04.005. URL `http://www.sciencedirect.com/science/article/pii/S1742287619301549`.

[224] Suneeta Satpathy, Sateesh K Pradhan, and Subhasish Mohapatra. Internet Usage Analysis Using Karl Pearson's Coefficient of Correlation-A Computer Forensic Investigation. *International Journal of Science and Research*, 3(11):2791–2794, 2014. ISSN 2319-7064.

[225] Mark Scanlon. Battling the digital forensic backlog through data deduplication. In *Innovative Computing Technology (INTECH), 2016 Sixth International Conference on*, pages 10–14. IEEE, 2016.

[226] Elizabeth D Schafer. Ancient science and forensics. *Forensic Science*, 1:41–45, 2008.

[227] Andreas Schmeling, Walter Reisinger, Gunther Geserick, and Andreas Olze. Age estimation of unaccompanied minors: Part I. General considerations. *Forensic science international*, 159:S61–S64, 2006.

[228] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[229] Hakan Sevimli, Ersin Esen, Tuğrul K Ateş, Ezgi C Ozan, Mashar Tekin, K Berker Loğoğlu, Ayça Müge Sevinç, Ahmet Saracoğlu, Adnan Yazici, and A Aydin Alatan. Adult image content classification using global features and skin region detection. In *Computer and Information Sciences*, pages 253–258. Springer, 2011.

[230] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[231] Nancy Signorielli. Children and Adolescents on Television: A Consistent Pattern of Devaluation. *The Journal of Early Adolescence*, 7(3):255–268, 1987.

[232] Aaron Smith. More than half of us adults trust law enforcement to use facial recognition responsibly. *Pew Research Center*, 2019.

[233] Leslie N Smith. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[234] Yu Su, S. Shan, X. Chen, and W. Gao. Hierarchical ensemble of gabor fisher classifier for face recognition. *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 6 pp.–96, 2006.

[235] James S Supancic and Deva Ramanan. Self-paced learning for long-term tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2379–2386, 2013.

[236] Avinash Swaminathan, Mridul Chaba, Deepak Kumar Sharma, and Yogesh Chaba. Gender Classification using Facial Embeddings: A Novel Approach. *Procedia Computer Science*, 167:2634–2642, 2020.

[237] Philipp Terhörst, Marco Huber, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Multi-algorithmic fusion for reliable age and gender estimation from face images. In *2019 22th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2019.

[238] Terminology and Semantics Interagency Working Group on Sexual Exploitation of Children. *Terminology guidelines for the protection of children from sexual exploitation and sexual abuse*. ECPAT International, 2016. URL `http://luxembourgguidelines.org/`.

[239] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[240] Min Jen Tsai, Cheng Liang Lai, and Jung Liu. Camera/mobile phone source identification for digital forensics. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2:221–224, 2007. ISSN 15206149. doi: 10.1109/ICASSP.2007.366212.

[241] G. Tsakalidis and K. Vergidis. A Systematic Approach Toward Description and Classification of Cybercrime Incidents. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(4):710–729, 2019. doi: 10.1109/TSMC.2017.2700495.

[242] Adrian Ulges and Armin Stahl. Automatic detection of child pornography using color visual words. In *2011 IEEE international conference on multimedia and expo*, pages 1–6. IEEE, 2011.

[243] United Nations General Assembly. *The UN Convention on the Rights of the Child*. UN Office of the High Commissioner for Human Rights, 1989.

[244] H.M.A. van Beek, E.J. van Eijk, R.B. van Baar, M. Ugen, J.N.C. Bodde, and A.J. Siemelink. Digital Forensics as a Service: Game on. *Digital Investigation*, 15:20 – 38, 2015. ISSN 1742-2876. doi: https://doi.org/10.1016/j.diin.2015.07.004. URL `http://www.sciencedirect.com/science/article/pii/S1742287615000857`. Special Issue: Big Data and Intelligent Data Analysis.

[245] Jenny Vestlund, Linda Langeborg, Patrik Sörqvist, and Mårten Eriksson. Experts on age estimation. *Scandinavian Journal of Psychology*, 50(4):301–307, 2009.

[246] S Vijayarani and M Vinupriya. Performance analysis of canny and sobel edge detection algorithms in image mining. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(8):1760–1767, 2013.

[247] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[248] Manuel C Voelkle, Natalie C Ebner, Ulman Lindenberger, and Michaela Riediger. Let me guess how old you are: effects of age, gender, and facial expression on perceptions of age. *Psychology and Aging*, 27(2):265, 2012.

[249] Ian Walden. *Computer crimes and digital investigations*. Oxford University Press, Inc., 2007.

[250] David Wall. *Cybercrime: The transformation of crime in the information age*, volume 4. Polity, 2007.

[251] David S Wall. Cybercrimes and the Internet. *Crime and the Internet*, pages 1–17, 2001.

[252] David S Wall. Dis-organised crime: Towards a distributed model of the organization of cybercrime. *The European Review of Organised Crime*, 2(2), 2015.

[253] Frank Wallhoff. Facial Expressions and Emotions Database, 2006. URL `http://www-prima.inrialpes.fr/FGnet/html/home.html`.

[254] J. Wang, J. Li, W. Yau, and E. Sung. Boosting dense SIFT descriptors and shape contexts of face images for gender recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 96–102, 2010. doi: 10.1109/CVPRW.2010.5543238.

[255] Sun-Chong Wang. Artificial neural network. In *Interdisciplinary computing in java programming*, pages 81–100. Springer, 2003.

[256] X. Wang, V. Ly, G. Lu, and C. Kambhamettu. Can We Minimize the Influence Due to Gender and Race in Age Estimation? In *2013 12th International Conference on Machine Learning and Applications*, volume 2, pages 309–314, Dec 2013. doi: 10.1109/ICMLA.2013.141.

[257] Xiaolong Wang and Chandra Kambhamettu. Age estimation via unsupervised neural networks. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6. IEEE, 2015.

[258] Heidi Weber, Antonio Cruz Rodriguez, and Américo Mateus. Emotion and mood in Design Thinking. *Design Doctoral Conference'16: TRANSversality - Proceedings of the DDC 3rd Conference*, 2016.

[259] Melissa Wells, David Finkelhor, Janis Wolak, and Kimberly J. Mitchell. Defining Child Pornography: Law Enforcement Dilemmas in Investigations of Internet Child Pornography Possession. *Police Practice and Research*, 8(3):269–282, 2007. doi: 10.1080/15614260701450765. URL `https://doi.org/10.1080/15614260701450765`.

[260] Gavin Willner and Paul Rowe. Alcohol servers' estimates of young people's ages. *Drugs: education, prevention and policy*, 8(4):375–383, 2001.

[261] Janis Wolak and Kimberly J Mitchell. Work exposure to child pornography in ICAC task forces and affiliates. *Retrieved from Crimes against Children Research Center*, 2009. URL `http://www.unh.edu/ccrc/pdf/Law%20Enforcement%20Work%20Exposure%20to%20CP.pdf`.

[262] Tao Wu, Pavan Turaga, and Rama Chellappa. Age estimation and face verification across aging using landmarks. *IEEE Transactions on Information Forensics and Security*, 7(6):1780–1788, 2012.

[263] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[264] Shuicheng Yan, Huan Wang, Thomas S Huang, Qiong Yang, and Xiaoou Tang. Ranking with uncertain labels. In *2007 IEEE International Conference on Multimedia and Expo*, pages 96–99. IEEE, 2007.

[265] Shuicheng Yan, Ming Liu, and Thomas S Huang. Extracting age information from local spatially flexible patches. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 737–740. IEEE, 2008.

[266] Colin Yao. Skin tone estimation and segmentation in MATLAB and OpenCV. `https://github.com/colin-yao/simple-skin-detection`, 2018.

[267] B. Yoo, Y. Kwak, Y. Kim, C. Choi, and J. Kim. Deep Facial Age Estimation Using Conditional Multitask Learning With Weak Label Expansion. *IEEE Signal Processing Letters*, 25(6):808–812, June 2018. ISSN 1558-2361. doi: 10.1109/LSP.2018.2822241.

[268] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[269] Ying Zhang, Jonathan Goh, Lei Lei Win, and Vrizlynn LL Thing. Image Region Forgery Detection: A Deep Learning Approach. In *SG-CRC*, pages 1–11, 2016.

[270] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

[271] Q. Zhao, J. Dong, H. Yu, and S. Chen. Distilling Ordinal Relation and Dark Knowledge for Facial Age Estimation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2020. doi: 10.1109/TNNLS.2020.3009523.

[272] Shaohua Kevin Zhou, Bogdan Georgescu, Xiang Sean Zhou, and Dorin Comaniciu. Image based regression using boosting method. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 541–548. IEEE, 2005.

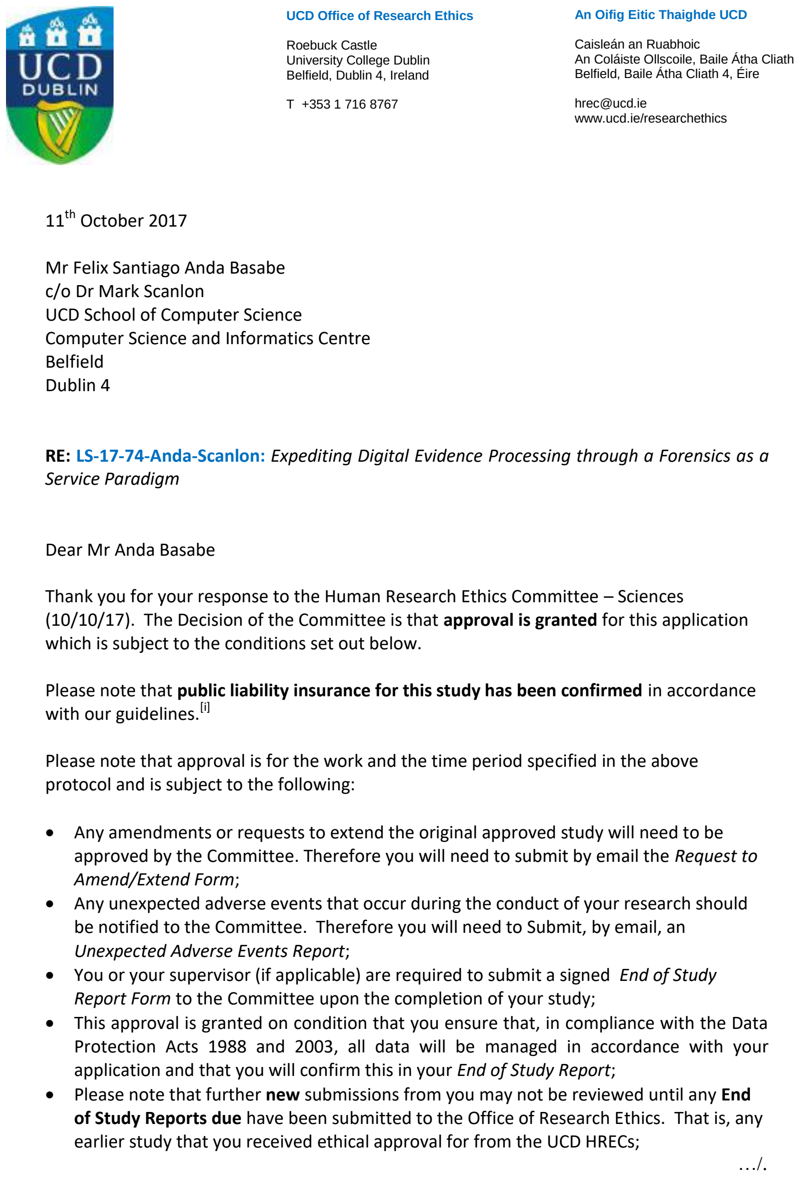
# Appendices

APPENDIX

# ONE

# ETHICAL APPROVALS AND EXEMPTIONS

# A.1 LS-17-74-Anda-Scanlon



UCD Office of Research Ethics

Roebuck Castle
University College Dublin
Belfield, Dublin 4, Ireland

T +353 1 716 8767

An Oifig Eitic Thaighde UCD

Caisleán an Ruabhoic
An Coláiste Ollscoile, Baile Átha Cliath
Belfield, Baile Átha Cliath 4, Éire

hrec@ucd.ie
www.ucd.ie/researchethics

11th October 2017

Mr Felix Santiago Anda Basabe
c/o Dr Mark Scanlon
UCD School of Computer Science
Computer Science and Informatics Centre
Belfield
Dublin 4

**RE:** **LS-17-74-Anda-Scanlon:** *Expediting Digital Evidence Processing through a Forensics as a Service Paradigm*

Dear Mr Anda Basabe

Thank you for your response to the Human Research Ethics Committee – Sciences (10/10/17). The Decision of the Committee is that **approval is granted** for this application which is subject to the conditions set out below.

Please note that **public liability insurance for this study has been confirmed** in accordance with our guidelines.[i]

Please note that approval is for the work and the time period specified in the above protocol and is subject to the following:

- Any amendments or requests to extend the original approved study will need to be approved by the Committee. Therefore you will need to submit by email the *Request to Amend/Extend Form*;
- Any unexpected adverse events that occur during the conduct of your research should be notified to the Committee. Therefore you will need to Submit, by email, an *Unexpected Adverse Events Report*;
- You or your supervisor (if applicable) are required to submit a signed *End of Study Report Form* to the Committee upon the completion of your study;
- This approval is granted on condition that you ensure that, in compliance with the Data Protection Acts 1988 and 2003, all data will be managed in accordance with your application and that you will confirm this in your *End of Study Report*;
- Please note that further **new** submissions from you may not be reviewed until any **End of Study Reports due** have been submitted to the Office of Research Ethics. That is, any earlier study that you received ethical approval for from the UCD HRECs;

…/.

- You may require copies of submitted documentation relating to this approved application and therefore we advise that you retain copies for your own records;
- Please note that the granting of this ethical approval is premised on the assumption that the research will be carried out within the limits of the law;
- Please also note that approved applications and any subsequent amendments are subject to a Research Ethics Compliance Review.

The Committee wishes you well with your research and look forward to receiving your End of Study Report. All forms are available on the website www.ucd.ie/researchethics please ensure that you submit the latest version of the relevant form. If you have any queries regarding the above please contact the Office of Research Ethics and please quote your reference in all correspondence.

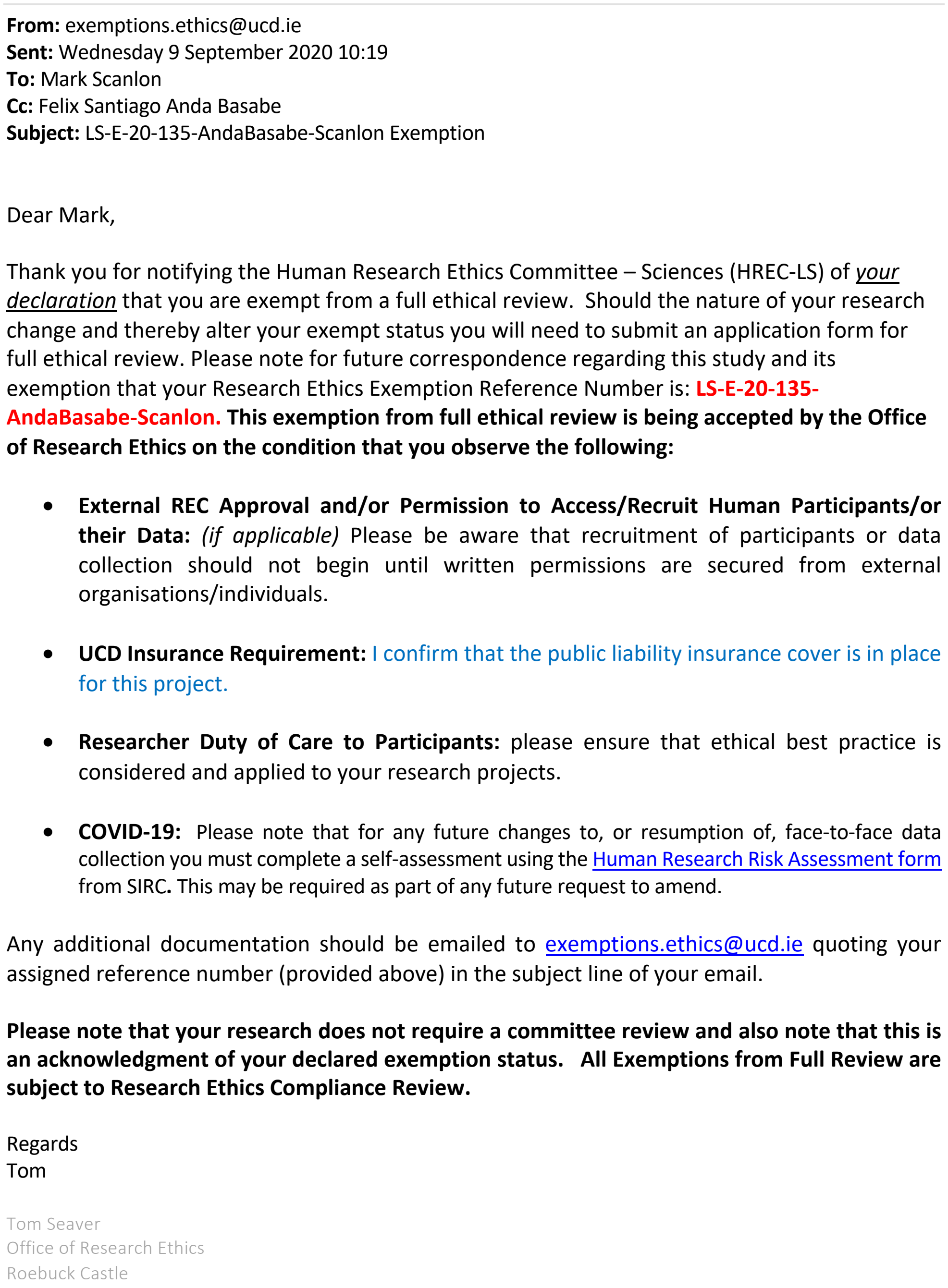
Yours sincerely,

John O'Dowd

Mr T. John O'Dowd
Chairman, Human Research Ethics Committee - Sciences

[i] http://www.ucd.ie/researchethics/information_for_researchers/insurance/

## A.2 LS-E-20-135-AndaBasabe-Scanlon



**From:** exemptions.ethics@ucd.ie
**Sent:** Wednesday 9 September 2020 10:19
**To:** Mark Scanlon
**Cc:** Felix Santiago Anda Basabe
**Subject:** LS-E-20-135-AndaBasabe-Scanlon Exemption

Dear Mark,

Thank you for notifying the Human Research Ethics Committee – Sciences (HREC-LS) of *your declaration* that you are exempt from a full ethical review. Should the nature of your research change and thereby alter your exempt status you will need to submit an application form for full ethical review. Please note for future correspondence regarding this study and its exemption that your Research Ethics Exemption Reference Number is: **LS-E-20-135-AndaBasabe-Scanlon. This exemption from full ethical review is being accepted by the Office of Research Ethics on the condition that you observe the following:**

- **External REC Approval and/or Permission to Access/Recruit Human Participants/or their Data:** *(if applicable)* Please be aware that recruitment of participants or data collection should not begin until written permissions are secured from external organisations/individuals.

- **UCD Insurance Requirement:** I confirm that the public liability insurance cover is in place for this project.

- **Researcher Duty of Care to Participants:** please ensure that ethical best practice is considered and applied to your research projects.

- **COVID-19:** Please note that for any future changes to, or resumption of, face-to-face data collection you must complete a self-assessment using the Human Research Risk Assessment form from SIRC**.** This may be required as part of any future request to amend.

Any additional documentation should be emailed to exemptions.ethics@ucd.ie quoting your assigned reference number (provided above) in the subject line of your email.

**Please note that your research does not require a committee review and also note that this is an acknowledgment of your declared exemption status. All Exemptions from Full Review are subject to Research Ethics Compliance Review.**

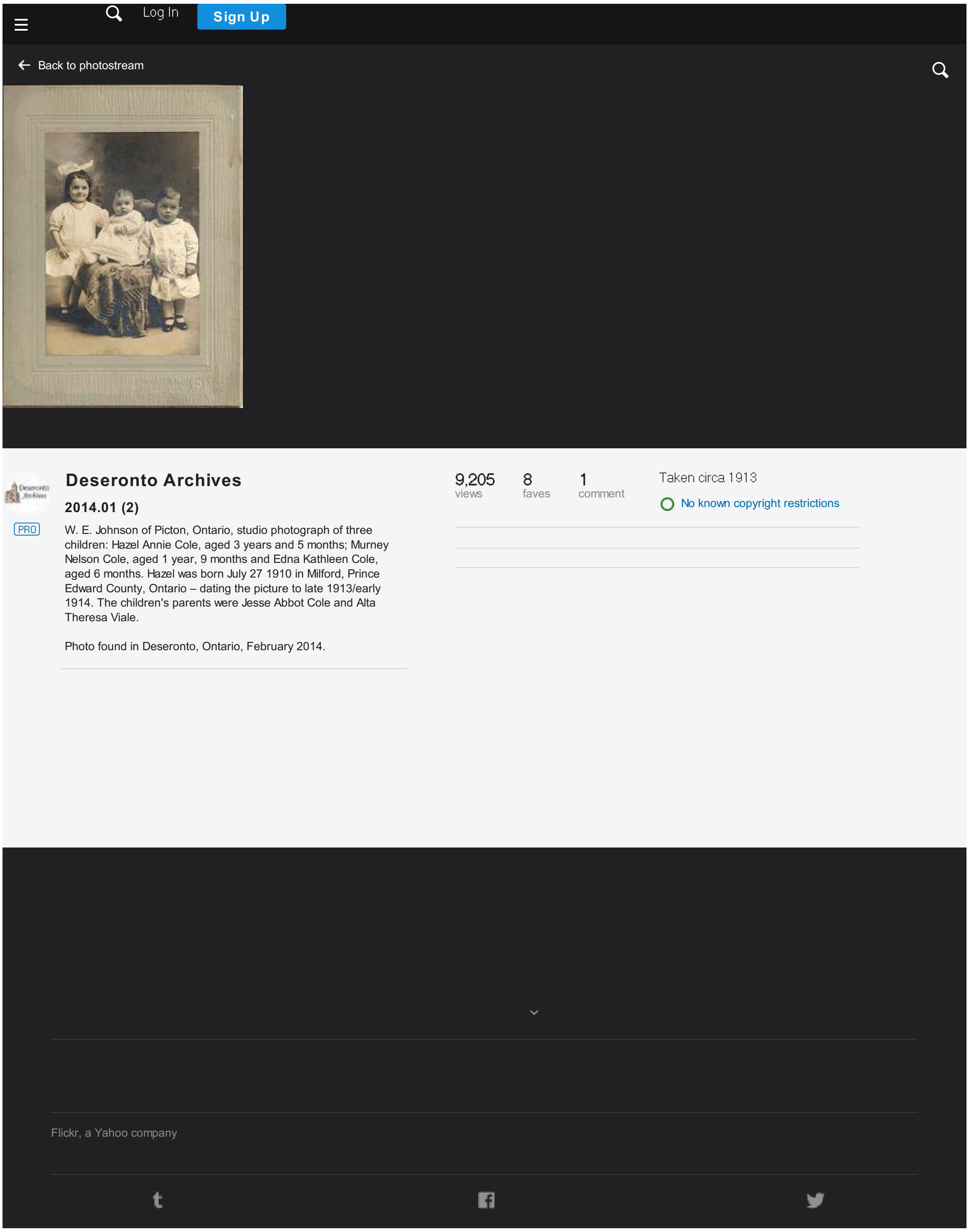
Regards
Tom

Tom Seaver
Office of Research Ethics
Roebuck Castle

# APPENDIX
# **TWO**

# VISAGE

# B.1 Audit Mechanism Document



**Deseronto Archives**

PRO

**2014.01 (2)**

W. E. Johnson of Picton, Ontario, studio photograph of three children: Hazel Annie Cole, aged 3 years and 5 months; Murney Nelson Cole, aged 1 year, 9 months and Edna Kathleen Cole, aged 6 months. Hazel was born July 27 1910 in Milford, Prince Edward County, Ontario – dating the picture to late 1913/early 1914. The children's parents were Jesse Abbot Cole and Alta Theresa Viale.

Photo found in Deseronto, Ontario, February 2014.

9,205 views 8 faves 1 comment

Taken circa 1913

○ No known copyright restrictions

Flickr, a Yahoo company

# B.2 Dataset Collection System User Manual

UCD Forensics and<br>Security Research Group

VISAGE

## Contents

1. Overview .......... 2
2. Accessing VISAGE .......... 2
3. Users .......... 3
    a. Registering a User .......... 3
    b. Signing in .......... 3
4. Campaigns .......... 4
    a. Starting a Campaign .......... 4
5. Voting .......... 6
    a. Mechanism .......... 6
    b. Submitting vote .......... 6
6. Statistics .......... 6
    a. General .......... 6
    b. Age Range .......... 7
    c. Users .......... 8
        i. Campaign details .......... 8
7. Positive Votes .......... 8
    a. Downloading Images and Metadata .......... 9

Figure 1 Welcome Screen .......... 2
Figure 2 Register User .......... 3
Figure 3 Registration Successful .......... 3
Figure 4 Home Screen .......... 4
Figure 5 Album details .......... 5
Figure 6 Votes .......... 6
Figure 7 General statistics .......... 7
Figure 8 Age Range Statistics .......... 7
Figure 9 Summary contributions .......... 8
Figure 10 Campaign Details .......... 8
Figure 11 Positive votes for 1 year old males .......... 9
Figure 12 JSON file .......... 9

UCD Forensics and
Security Research Group

1. Overview

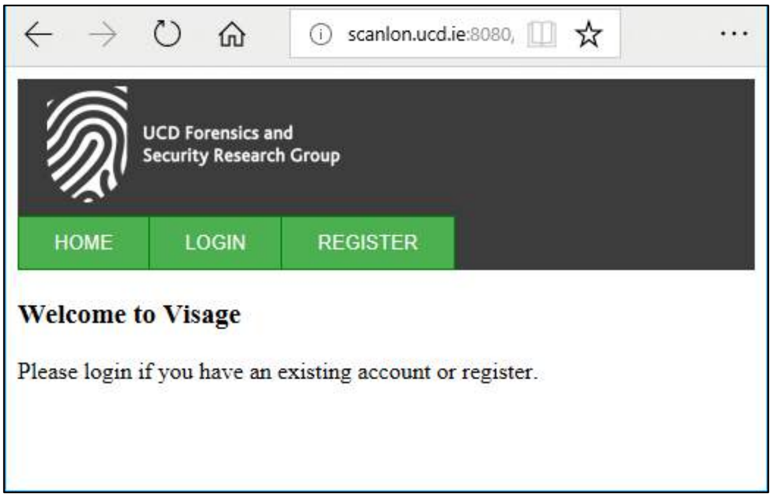
VisAGe is a web-based voting system aimed to label facial images with age and gender. The system has been built with a dataset of images that have been obtained from Flickr. The hybrid interaction with both humans and machines to tag data is an approach that can lessen the effort of users in order to contribute to research. The users can vote from a digital handheld device, laptop, computer, etc. Only a web browser is required. Once three votes are committed, the image becomes part of a positive voting dataset available for researchers. Moreover, different campaigns are created so the voting can be sorted by age, gender and type of faces such as single, multiple and others (No face). Finally, images are downloadable with the metadata corresponding to the Microsoft Azure Face API results. The results are presented in a JSON format and are attached to the files in a zip format.

2. Accessing VISAGE

To access the VisAGe web application, please use a browser and type:

http://scanlon.ucd.ie:8080/gallery

The following screen should be displayed:



Welcome to Visage

Please login if you have an existing account or register.

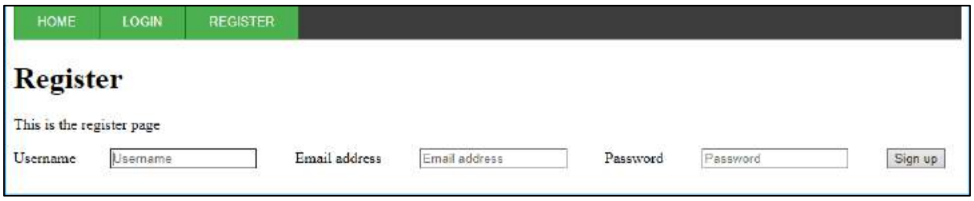
*Figure 1 Welcome Screen*

3. Users

In order to access Visage, a valid user must be registered. The system will allow the freely creation of users and sensitive data is stored on the database with modern password hashing.

a. Registering a User

On the web application, select REGISTER on the menu. Type a new valid username, email address, password.

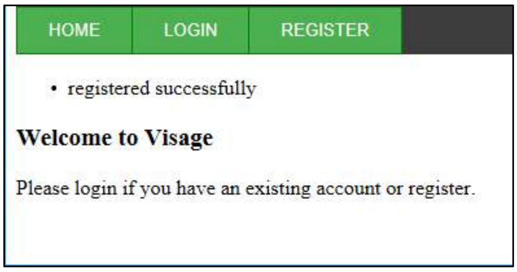


Register

This is the register page

*Figure 2 Register User*

After pressing Sign up, the following screen should be displayed:



- registered successfully

Welcome to Visage

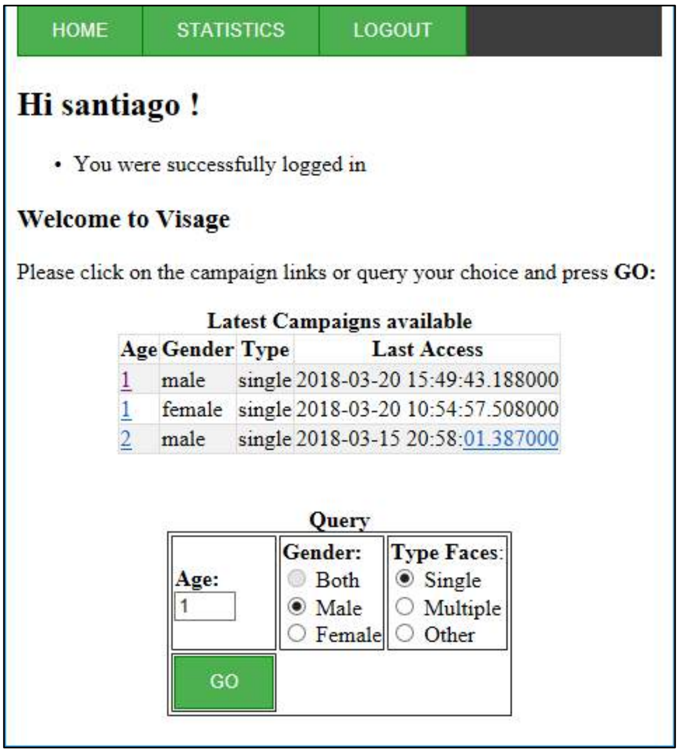
Please login if you have an existing account or register.

*Figure 3 Registration Successful*

b. Signing in

Once a user has been created, you should be able to login with the credentials provided. If you are having trouble signing in, please contact the system administrator.

If the login was successful, the screen depicted in Figure 4 Home Screen should be displayed:



Hi santiago !

- You were successfully logged in

Welcome to Visage

Please click on the campaign links or query your choice and press **GO:**

Latest Campaigns available

| Age | Gender | Type | Last Access |
|---|---|---|---|
| 1 | male | single | 2018-03-20 15:49:43.188000 |
| 1 | female | single | 2018-03-20 10:54:57.508000 |
| 2 | male | single | 2018-03-15 20:58:01.387000 |

Query

| Age: | Gender: | Type Faces: |
|---|---|---|
| 1 | ○ Both<br>● Male<br>○ Female | ● Single<br>○ Multiple<br>○ Other |

*Figure 4 Home Screen*

As seen in Figure 4 Home Screen, we have a menu where we can select HOME, STATISTICS or LOGOUT.

After the welcoming message, the top 5 latest campaigns will be displayed and a Query functionality where we can select a personalized query.

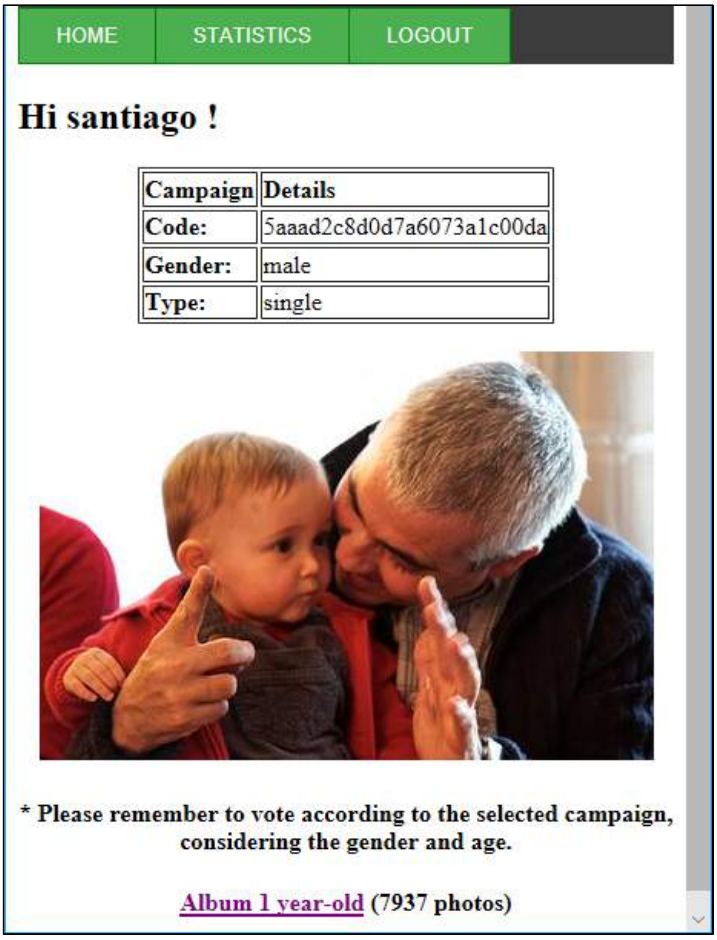
4. Campaigns

The system is designed to work with campaigns, so we can track the progress by age and gender.

a. Starting a Campaign

As seen in Figure 4 Home Screen, you can either select a campaign displayed in the top 5 latest campaigns by clicking the link over the values of the age in the age column. The other option would be to select a personalized campaign by typing in the age, selecting the gender, choosing the type of face and finally clicking on the GO button.

The following screen should be displayed as shown in Figure 5 Album details:



Hi santiago !

| Campaign | Details |
|---|---|
| Code: | 5aaad2c8d0d7a6073a1c00da |
| Gender: | male |
| Type: | single |

* Please remember to vote according to the selected campaign, considering the gender and age.

Album 1 year-old (7937 photos)

*Figure 5 Album details*

Details of the campaign are shown such as Code, Gender, Type and the number of images available for voting.

5. Voting
   a. Mechanism

When 3 votes are submitted, the system stops generating the voted images and provides new images.

   b. Submitting vote

The process of submitting a vote is done after discarding or approving images that are relevant to the campaign selected.

Figure 6 Votes below depicts how images are voted positively and negatively.

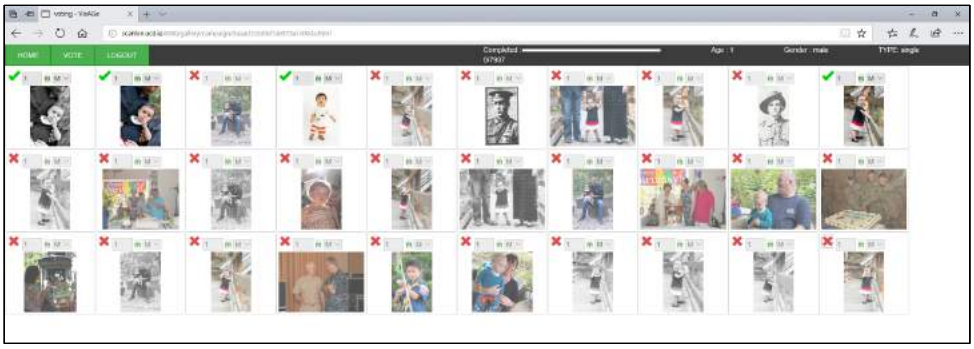


*Figure 6 Votes*

Once the images are selected, the vote button must be pressed. The following text should appear on the top side of the web page:

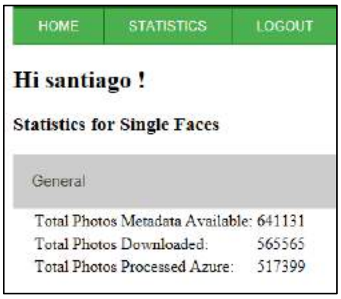
*"Thank you for your vote! next set of images are presented".*

6. Statistics

In order to track the work done, a statistics option is available. There are 3 sections presented: General, Age Range and Users

   a. General

This panel gives us an overview of the total of metadata, photos created, and photos processed by Microsoft Azure as shown in Figure 7 General statistics.



**Hi santiago !**

**Statistics for Single Faces**

General

| | |
|---|---|
| Total Photos Metadata Available: | 641131 |
| Total Photos Downloaded: | 565565 |
| Total Photos Processed Azure: | 517399 |

*Figure 7 General statistics*

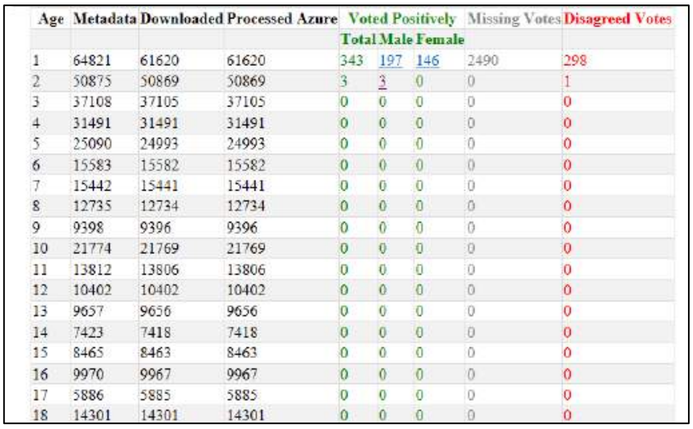
b. Age Range

This panel has information by age and calculates the total. The descriptions are depicted in Table 1 Column description.

| Column | Description |
|---|---|
| Voted Positively | Image has been voted positively 3 times by different users |
| Missing Votes | Votes are pending on different Users in order to gain 3 votes |
| Disagreed Votes | Votes are different by different users |

*Table 1 Column description*

In the panel shown in Figure 8 Age Range Statistics, links can be noticed over positive votes

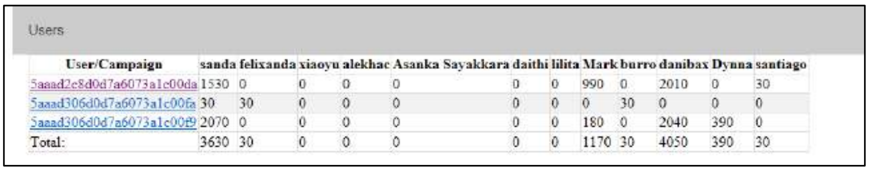


| Age | Metadata | Downloaded | Processed Azure | Voted Positively | | | Missing Votes | Disagreed Votes |
|---|---|---|---|---|---|---|---|---|
| | | | | Total | Male | Female | | |
| 1 | 64821 | 61620 | 61620 | 343 | 197 | 146 | 2490 | 298 |
| 2 | 50875 | 50869 | 50869 | 3 | 3 | 0 | 0 | 1 |
| 3 | 37108 | 37105 | 37105 | 0 | 0 | 0 | 0 | 0 |
| 4 | 31491 | 31491 | 31491 | 0 | 0 | 0 | 0 | 0 |
| 5 | 25090 | 24993 | 24993 | 0 | 0 | 0 | 0 | 0 |
| 6 | 15583 | 15582 | 15582 | 0 | 0 | 0 | 0 | 0 |
| 7 | 15442 | 15441 | 15441 | 0 | 0 | 0 | 0 | 0 |
| 8 | 12735 | 12734 | 12734 | 0 | 0 | 0 | 0 | 0 |
| 9 | 9398 | 9396 | 9396 | 0 | 0 | 0 | 0 | 0 |
| 10 | 21774 | 21769 | 21769 | 0 | 0 | 0 | 0 | 0 |
| 11 | 13812 | 13806 | 13806 | 0 | 0 | 0 | 0 | 0 |
| 12 | 10402 | 10402 | 10402 | 0 | 0 | 0 | 0 | 0 |
| 13 | 9657 | 9656 | 9656 | 0 | 0 | 0 | 0 | 0 |
| 14 | 7423 | 7418 | 7418 | 0 | 0 | 0 | 0 | 0 |
| 15 | 8465 | 8463 | 8463 | 0 | 0 | 0 | 0 | 0 |
| 16 | 9970 | 9967 | 9967 | 0 | 0 | 0 | 0 | 0 |
| 17 | 5886 | 5885 | 5885 | 0 | 0 | 0 | 0 | 0 |
| 18 | 14301 | 14301 | 14301 | 0 | 0 | 0 | 0 | 0 |

*Figure 8 Age Range Statistics*

c. Users

In the user's panel, there is a summary of contributions by user by campaign observed in Figure 9 Summary contributions.



Users

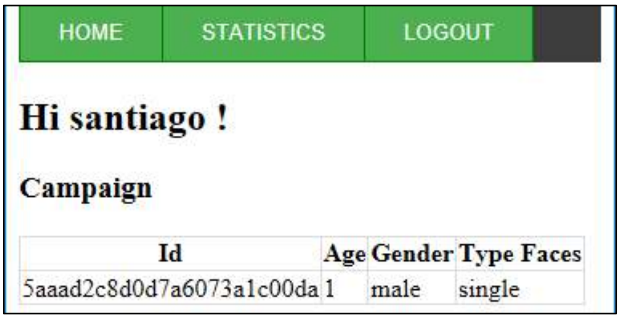
| User/Campaign | sanda | felixanda | xiaoyu | alekhac | Asanka Sayakkara | daithi | lilita | Mark | burro | danibax | Dynna | santiago |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5aaad2c8d0d7a6073a1c00da | 1530 | 0 | 0 | 0 | 0 | 0 | 0 | 990 | 0 | 2010 | 0 | 30 |
| 5aaad306d0d7a6073a1c00fa | 30 | 30 | 0 | 0 | 0 | 0 | 0 | 0 | 30 | 0 | 0 | 0 |
| 5aaad306d0d7a6073a1c00f9 | 2070 | 0 | 0 | 0 | 0 | 0 | 0 | 180 | 0 | 2040 | 390 | 0 |
| Total: | 3630 | 30 | 0 | 0 | 0 | 0 | 0 | 1170 | 30 | 4050 | 390 | 30 |

*Figure 9 Summary contributions*

Furthermore, the user can access the campaign details by clicking on the campaign id.

i. Campaign details

The details of a specific campaign can be seen in the STATISTICS option, on the Users panel by clicking on the code of the campaign:



Hi santiago !

Campaign

| Id | Age | Gender | Type Faces |
|---|---|---|---|
| 5aaad2c8d0d7a6073a1c00da | 1 | male | single |

*Figure 10 Campaign Details*

7. Positive Votes

The images that are the outcome of 3 votes can be shown by clicking on the links discussed in the section 7.

A web page is displayed with the positively voted images:

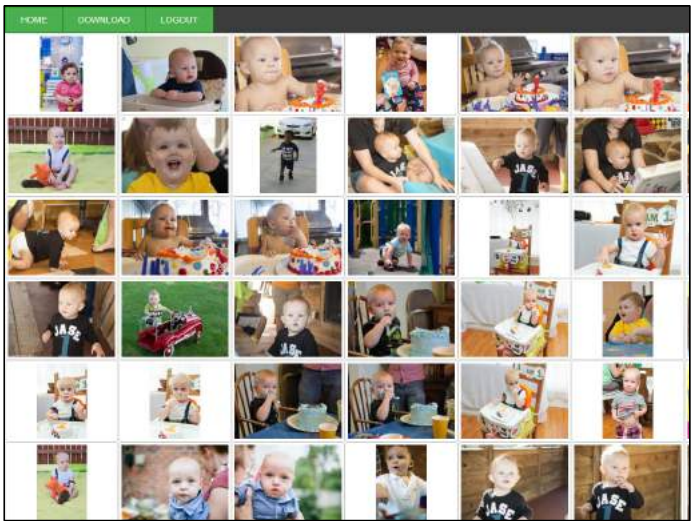
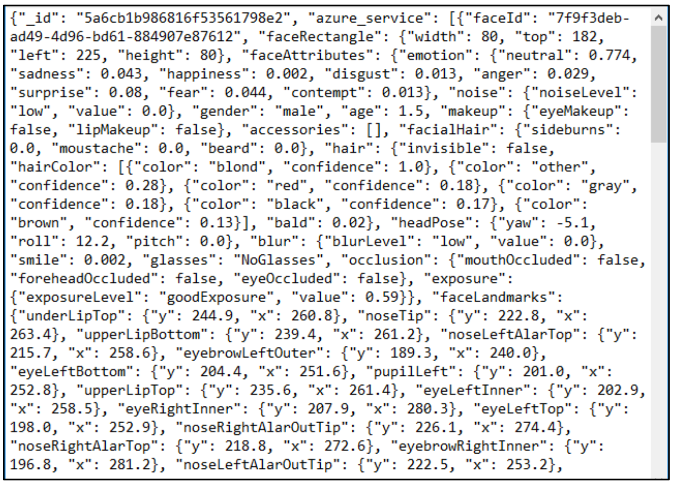


*Figure 11 Positive votes for 1 year old males*

a. Downloading Images and Metadata

By clicking the download button, a zip file is downloaded containing images and a JSON file with the Azure Face API metadata as shown in Figure 12.



```
{"_id": "5a6cb1b986816f53561798e2", "azure_service": [{"faceId": "7f9f3deb-
ad49-4d96-bd61-884907e87612", "faceRectangle": {"width": 80, "top": 182,
"left": 225, "height": 80}, "faceAttributes": {"emotion": {"neutral": 0.774,
"sadness": 0.043, "happiness": 0.002, "disgust": 0.013, "anger": 0.029,
"surprise": 0.08, "fear": 0.044, "contempt": 0.013}, "noise": {"noiseLevel":
"low", "value": 0.0}, "gender": "male", "age": 1.5, "makeup": {"eyeMakeup":
false, "lipMakeup": false}, "accessories": [], "facialHair": {"sideburns":
0.0, "moustache": 0.0, "beard": 0.0}, "hair": {"invisible": false,
"hairColor": [{"color": "blond", "confidence": 1.0}, {"color": "other",
"confidence": 0.28}, {"color": "red", "confidence": 0.18}, {"color": "gray",
"confidence": 0.18}, {"color": "black", "confidence": 0.17}, {"color":
"brown", "confidence": 0.13}], "bald": 0.02}, "headPose": {"yaw": -5.1,
"roll": 12.2, "pitch": 0.0}, "blur": {"blurLevel": "low", "value": 0.0},
"smile": 0.002, "glasses": "NoGlasses", "occlusion": {"mouthOccluded": false,
"foreheadOccluded": false, "eyeOccluded": false}, "exposure":
{"exposureLevel": "goodExposure", "value": 0.59}}, "faceLandmarks":
{"underLipTop": {"y": 244.9, "x": 260.8}, "noseTip": {"y": 222.8, "x":
263.4}, "upperLipBottom": {"y": 239.4, "x": 261.2}, "noseLeftAlarTop": {"y":
215.7, "x": 258.6}, "eyebrowLeftOuter": {"y": 189.3, "x": 240.0},
"eyeLeftBottom": {"y": 204.4, "x": 251.6}, "pupilLeft": {"y": 201.0, "x":
252.8}, "upperLipTop": {"y": 235.6, "x": 261.4}, "eyeLeftInner": {"y": 202.9,
"x": 258.5}, "eyeRightInner": {"y": 207.9, "x": 280.3}, "eyeLeftTop": {"y":
198.0, "x": 252.9}, "noseRightAlarOutTip": {"y": 226.1, "x": 274.4},
"noseRightAlarTop": {"y": 218.8, "x": 272.6}, "eyebrowRightInner": {"y":
196.8, "x": 281.2}, "noseLeftAlarOutTip": {"y": 222.5, "x": 253.2},
```

*Figure 12 JSON file*