Improving Borderline Adulthood Facial Age Estimation through Ensemble Learning

Felix Anda*
felix.anda@ucdconnect.ie
University College Dublin
Dublin, Ireland

David Lillis*
david.lillis@ucd.ie
University College Dublin
Dublin, Ireland

Aikaterini Kanta*
aikaterini.kanta@ucdconnect.ie
University College Dublin
Dublin, Ireland

Brett A. Becker†
brett.becker@ucd.ie
University College Dublin
Dublin, Ireland

Elias Bou-Harb†
ebouharb@fau.edu
Florida Atlantic University
Boca Raton, FL, USA

Nhien-An Le-Khac*
an.lekhac@ucd.ie
University College Dublin
Dublin, Ireland

Mark Scanlon†
mark.scanlon@ucd.ie
University College Dublin
Dublin, Ireland

ABSTRACT

Achieving high performance for facial age estimation with subjects in the borderline between adulthood and non-adulthood has always been a challenge. Several studies have used different approaches from the age of a baby to an elder adult and different datasets have been employed to measure the mean absolute error (MAE) ranging between 1.47 to 8 years. The weakness of the algorithms specifically in the borderline has been a motivation for this paper. In our approach, we have developed an ensemble technique that improves the accuracy of underage estimation in conjunction with our deep learning model (DS13K) that has been fine-tuned on the Deep Expectation (DEX) model. We have achieved an accuracy of 68% for the age group 16 to 17 years old, which is 4 times better than the DEX accuracy for such age range. We also present an evaluation of existing cloud-based and offline facial age prediction services, such as Amazon Rekognition, Microsoft Azure Cognitive Services, How-Old.net and DEX.

CCS CONCEPTS

• **Applied computing** → **Investigation techniques**; **Evidence collection, storage and analysis**; • **Computing methodologies** → *Ensemble methods.*

KEYWORDS

Underage Photo Datasets, Deep Learning, Digital Forensics, Child Exploitation Investigation, Facial Recognition

1 INTRODUCTION

Automated facial age estimation is the application of non-manual processes to measure the age of a person by analysing specific facial features with the use of artificial intelligence. Facial age related products are becoming increasingly popular in our daily life. Teenagers and adults are using ageing filters for entertainment purposes that are available with Snapchat which have become viral over the past years. These new face-ageing techniques could boost search for wanted criminals or missing people. The usage of facial recognition has incremented exponentially. Furthermore, 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 is a widely-used technology that maps facial features from images to detect faces and recognise the associated identity. Applications have been commonly found in airports, mobile devices and certain web pages [12]. Entertainment venues, alcohol, tobacco and certain social media services require an age verification process. 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. Facial age detection could also be used to prevent unauthorised consumption or purchase of certain goods or services. Undocumented criminals are open to deceive authorities about their age to avoid the judicial system; however, an automated age detector could impede their attempt to bypass the system.

Accurate facial age estimation has long been a difficult task for both human experts and specialised machine learning algorithms. Moreover, the influence of factors, such as environment, health habits, lifestyle, makeup, emotions, and uncontrolled lightning hinder the age estimation process [15]. We have studied the possibility of including artificial intelligence as a means to detect and analyse evidence that may be presented in court. Specifically, we have focused on the improvement of facial age estimation algorithms for the identification of victims/suspects and its applications to child sexual exploitation material (CSEM) and child sexual abuse material (CSAM) investigations[1]. Challenges arise due to the factors previously mentioned, thus hampering age classification accuracy, especially for borderline cases between underage and adult subjects. 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 usage of Deep Learning in several fields has become the latest trend: at the end of 2018, Google introduced an AI tool (freely

*UCD Forensics and Security Research Group - https://www.forensicsandsecurity.com
†Cyber Threat Intelligence Laboratory

[1]These are the terms recommended by the Luxembourg Guidelines (http://luxembourgguidelines.org/)

available for non-governmental organisations and industry partners) to assist organisations in detecting and reporting child sexual abuse material online [13]. With the emergence of AI and its state-of-the-art branches including computer vision, machine learning and deep learning, age determination has improved significantly. Neural networks learn by processing thousands of images so that they can predict the age of future unseen images at an accuracy that surpasses human facial age perception capacities.

Given the quantity of digital content being created daily, the previous approach of manual evidence analysis is unfeasible [17]. However, machines require training to deliver accurate estimations. The training process demands a large volume of labelled data, extensive time, and computer resources to understand traits present in digital portraits. In a previous study, several age estimation services were evaluated throughout an age range of 0 to 77 years old [2]. With this study it was found that the real culprit of inaccurate age predictions for minors is linked to the lack of appropriate datasets with adequate age labels. Nonetheless, data collection of underage images is surrounded by ethical and moral concerns. Personal identifiable information, such as name, gender, age, and additional information must be handled with care and the exposure of sensitive information by uploading underage images in an unencrypted Internet can be detrimental. Conversely, data collection with the appropriate safeguards could assist missing children and detect previously unknown child abuse material.

Child exploitation investigations are one of the more common investigation types in digital forensic laboratories throughout the world [1]. These investigations have become an arduous task due to the increasing usage of anonymization tools, private P2P networks and cloud-based KVM systems [9]. Worldwide, law enforcement and child protection communities have been fighting to diminish CSEM and human trafficking. Automated age detection techniques can be used to reduce work exposure to incriminating archives of indecent images; therefore, reducing the psychological ramifications. Such techniques have also been exercised for image classification and categorisation according to age, gender, objects contained therein, and the location in which each image was taken, all of which are useful to CSEM investigators.

1.1 Contribution of this Work

The contribution of this work can be summarised as:

- Comprehensive performance evaluation of offline and cloud-based facial recognition models.
- The development and evaluation of a novel deep learning based underage subject classification model, `DS13K` with N=12792 images, 80% for training and 20% for testing.
- 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.

2 LITERATURE REVIEW/STATE OF THE ART

2.1 Automated Age Estimation

The human face can reveal important information, such as gender, approximate age, skin tone, 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. [6] 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. [16] 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 and the respective age group.

The mean absolute error (MAE) and the mean absolute error per age (MAE/A) are the performance metrics used throughout this paper. The former is the average difference between the predicted age and the ground truth; the latter is the MAE grouped by the age.

In the past two decades, error rates have decreased remarkably. A MAE of 1.47 was achieved by Ratnayake et al. [23] in 2014 by accomplishing an AdaBoost[2] fusion of several state-of-the-art classifiers (including Fisher's LDA, Neural Networks, and Support Vector Machine). Nevertheless, this study was executed over a limited private dataset of 50 female images with an age range from 10 to 19, which is indicative of the scarcity of suitable images of this type. In 2011, Luu et al. [20] were able to obtain a MAE of 4.1 (which has been typical of techniques utilising the FG-NET database). The *contourlet* appearance model used was more accurate and faster at localising facial landmarks than active appearance Models. Ferguson and Wilkinson [10] acknowledged poor accuracy results for age estimation on juvenile faces by human observation. Influence of age, sex and occupation is nullified in the outcome. Moreover, female age estimation was more accurate in younger age groups and male age prediction were more precise after 11 years of age.

2.2 Transfer Learning

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 [5].

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 pretrained 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. In our study, we are attempting to obtain a sufficient quantity of labelled facial age images; however issues arise due to copyright restrictions, GDPR, and ethical concerns. Therefore, a transfer learning solution is required. In further studies, Dong et al. [7] exploited the transfer learning strategy to train deep convolutional neural networks from pretrained models due to the scarcity of age labelled face images.

[2] AdaBoost is a machine learning boosting algorithm that iteratively builds an ensemble of models [25].

Transfer learning is usually expressed through the use of pretrained models, which are simply models created to solve a specific problem and are suitable for re-usability. Less training data is required when successfully transferring a pretrained model to another task.

2.3 Face Ageing Datasets

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 and inaccurate age labelling.

IMDB-WIKI is the largest public facial age computer annotated age and gender dataset [24] and has been subject of hundreds of facial recognition studies. The images were scraped from thousands of celebrities in IMDB[3] and correlated with Wikipedia[4]. The collection is quite considerable as the figures reach over half a million; 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.

The FG-NET [27] 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. Grd and Bača [14] produced a private database in 2016 called ageCFBP with a wider age range. In the same year, Boys2Men was released as another private database focused on male child images [3].

MEDS [11] is a mugshot dataset of male and female deceased subjects with the oldness feature annotated but does not contain images of underage individuals. The FERET dataset contains around 14,000 images and is pertinent to face detection [22]. The age labelling is based solely on human observation.

The OUI-Adience set is a public collection of labelled images obtained by online facial images of Flickr "in the wild". Although Eidinger et al. [8] has stated that they use Creative Commons license for their images, we have detected from a sample of 10,842 images, that 89.55% are associated to images with copyright; therefore, we have avoided the use of such dataset. Another dataset that uses Flickr as a source is the Yahoo Flickr Creative Commons 100M (YFCC100M) that was released in 2014 [26]. 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).

For our studies, a hybrid dataset was created from a variety of those available (IMDB, WIKI, FG-NET, MEDS) using the dataset generator software published by Anda et al. [2].

3 EXISTING TOOLS AND MODELS

In this section, the current tools for age estimation that are classified in two categories: Offline and Online. For the former, the tools are associated with pretrained models, where the architecture is known and the training dataset may or may not be shared. For the latter, the tools are hosted as cloud services, and the architecture of the neural network and the training dataset are generally unknown.

The main advantage of using an offline pretrained model is that they are usually shared by researchers either in frameworks, such as Caffe[5], Caffe2[6], Keras[7] or Pytorch[8] and thus have no cost. Nonetheless, online tools are associated with machine learning as a service and require a payment per transaction but are much easier to invoke; no installation is required and less local computational power is used.

The age and gender classification using Convolutional Neural Networks (CNNs) is an offline model that was built on the Adience dataset and released in 2005 [18]. This pretrained model consisted of a CNN architecture that was adapted to work even though the amount of learning data was scarce. Similarly, the ranking CNN for age estimation model was released in 2017 and is also an offline model that is available in the Model Zoo[9]. This model contains a series of basic CNNs that were fine-tuned from the base network trained on the Adience dataset. The result is a binary output and is ultimately added to the final prediction [4].

According to Economy Watch in 2010 [29], Amazon acquired "Rekognition" from an Artificial Intelligence start-up company from California called Orbeus. The company had developed a facial recognition software that detected traits on images with the use of a library based on Artificial Neural Networks which are computing systems that learn to accomplish tasks by observing examples rather than executing a specific algorithm and are structured by an initial input layer of neurons, one or more hidden layers, and a final layer of output neurons [28].

The Kairos service has been used for age prediction and face detection; however according to Anda et al. [2], the age estimation performance was lagging behind the rest of the classifiers included in that study. On the contrary, Microsoft Azure Machine Learning is a fully managed cloud service that is powered by a considerable number of machine learning algorithms aimed for scientists, data analysts and developers [21]. Per Weber et al. [30], it is suggested that Microsoft Azure Cognitive Service uses Multi-layered deep learning technology and is within the top performers for age estimation. Finally, DEX has been subject to hundreds of studies in fields, such as computer vision, deep learning face recognition and age estimation. The huge dataset of over half a million subjects has been used by several researchers and the model has been trained in multiple frameworks, such as Caffe and Keras[10].

Google has not yet released a fully-fledged age estimation service based on image analysis to the public. The Google Vision Cloud API includes facial recognition and facial landmark features, but only allow the recognition of subjects to be categorised as a minor or non-minor and safe search capabilities, such as the recognition of adult content. It could be suggested that the introduction of the Google tool to assist organisations in detecting and reporting child sexual abuse material online previously mentioned in Section 1,

[3] https://www.imdb.com/
[4] https://www.wikipedia.org/
[5] https://caffe.berkeleyvision.org/
[6] https://caffe2.ai/
[7] https://keras.io/
[8] https://pytorch.org/
[9] https://modelzoo.co/model/using-ranking-cnn-for-age-estimation
[10] https://github.com/yu4u/age-gender-estimation

is the combination of both the minor/non-minor detector and the adult content detector.

Finally, How-old.net is an application linked to the Microsoft cognitive services and part of Microsoft's *Project Oxford*. In recent years, the tool went viral on social media and was used mainly for entertainment. Today it can be used to predict underage images with a fairly high accuracy as shown in our study.

4 DATASET CURATION FOR PERFORMANCE EVALUATION

In order to perform unbiased experimentation with the four services identified, it was necessary to construct a balanced dataset. Thus, we ensured that there were an equal number of images collected for each age. The dataset generator proposed in [2] was used and additional modules for the datasets that are to be discussed in this section were implemented[11].

Because the focus of this paper is on the boundary between minority and adulthood, older ages were not considered. Thus, the dataset was limited to an age range of 0 to 25 inclusive. For this dataset, 492 images per age 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[12]. Only photos that were available under an appropriate Creative Commons 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 [31]. IMDB and WIKI photos were avoided but still used in a low proportion. This dataset was used for the experiment described in Section 5.1. Each image is a single frontal face that was cropped and aligned with DLIB[13] with a dimension of 200 x 200 pixels. Each image was processed by a face detector either by the DLIB libraries by using Histogram Oriented Gradients or Convolutional Neural Networks, or the face detection provided by each service discussed in Section 3. Initially we had a collection of 15,000 images but due to non-face recognition, the figure decreased and in order to maintain a balanced dataset, the images had to be reduced to 492 per class hence, we limited the dataset to a total size of 12,792.

5 EXPERIMENTS AND RESULTS

Three experiments were conducted and the MAE was calculated with the formula depicted in Equation 1, the results of which are presented in the subsections that follows. The first experiment, discussed in Section 5.1 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. In addition to the services, our deep learning model, `DS13K` was created. The second experiment involves the evaluation of DS13K. The model performance reached an accuracy of 55.38% placing it in the top 3 performers after Bagging Regressor and the Gradient Boosting Regressor. The model is described in Section 5.2. The final experiment introduces ensemble machine learning techniques to establish whether these will be useful tools to improve upon the performance of the four systems. This is presented in Section 5.3.

$$MAE = \sum_{i=1}^{n} \frac{|predicted_i - real_i|}{n} \quad (1)$$

5.1 Underage Range Estimation

The evaluation for the first experiment focused on samples from 0 to 25. The results of the evaluation are shown in Figure 1, with the average predicted age for each service plotted against the subjects' actual ages. The MAE for each service can be seen in Figure 2 and the average MAE for underage subjects is presented in Table 1.

From these figures, it can be seen that Amazon Rekognition 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, DEX's 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 a lack of sufficient sample images used to train the Deep Expectation model for very young subjects, and is the primary reason why DEX's 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 1. 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. AWS's 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
Amazon Rekognition	3.349
How-Old.net	5.281
Microsoft Azure	5.347
(D)eep (EX)pectation	6.936

Table 1: Mean Absolute Error for Underage Images per Service.

5.2 Development of a Deep Learning Model for Age Estimation (DS13K)

The previously-mentioned DEX model in Section 3 was built on a VGG-16 architecture. For the development of our model, transfer learning was used; our DS13K model was fine-tuned on DEX in order to take advantage of the preexisting layer weights. Furthermore, the 12,792 images used for training and testing (80% and

[11] https://bitbucket.org/4nd4/image_database

[12] Appropriate ethical approval was awarded for this data gathering process from our research institution (University College Dublin)

[13] C++ toolkit containing machine learning algorithms http://dlib.net/.

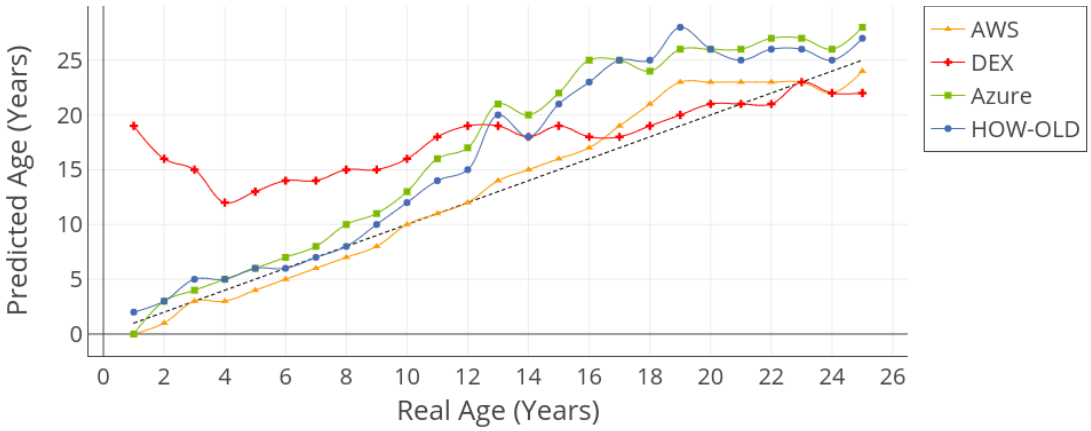

Figure 1: Average Estimated Age from each Service Compared with Actual Age.

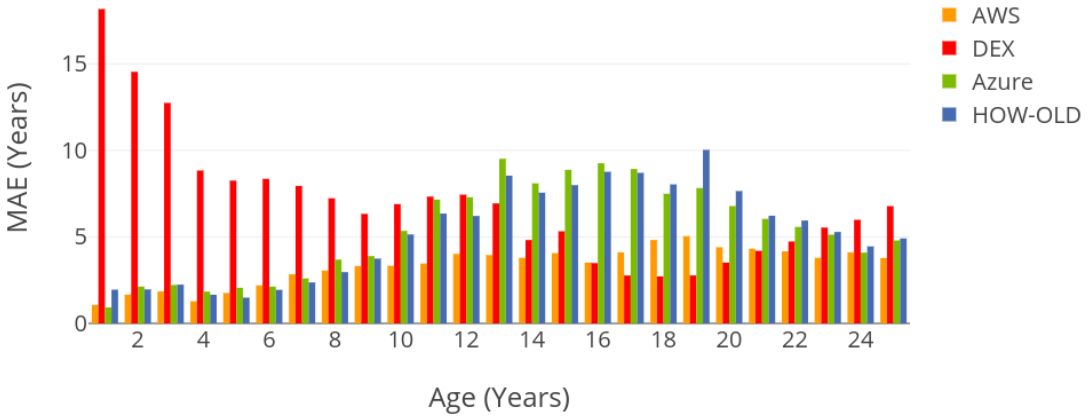

Figure 2: Mean Absolute Error per Age by Service.

20% respectively) came from sources described in Section 4. Each input image was resized to a dimension of 224 x 224 pixels and the output had a size of 5 (Multi-class classifier) and were mapped to a value pertaining to the following age range 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[14], which indicates that the classification of subjects into one of these age ranges is sufficient, and that precise age estimation is not crucial for investigators.

To supervise the input of the model, each age class was split into two and the average faces were calculated as depicted in figure 3. The accuracy per age group as well as the average accuracy per service is in Table 2, where the best-performing figure for each age range is illustrated in bold. DS13K has the best average performance followed closely by AWS. 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. The second-highest accuracy for this range was AWS with 15%. As illustrated

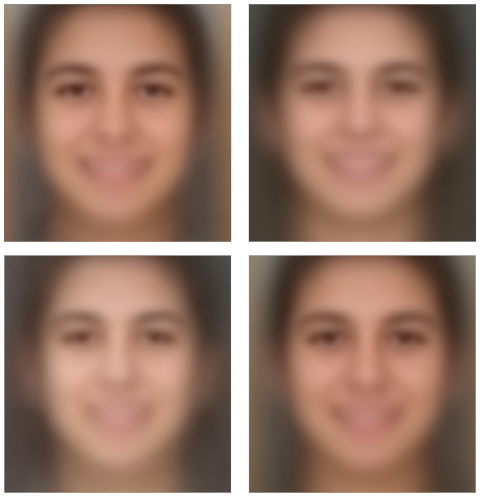

Figure 3: Average Faces of DS13K Subjects between 16 to 17 Years Old.

[14]https://www.europol.europa.eu/publications-documents/criminal-networks-involved-in-trafficking-and-exploitation-of-underage-victims-in-eu

previously in Figure 1, 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 (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 2: Accuracy per Group per Service.

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

5.3 Comparison with Ensemble Learning Techniques.

The third experiment was intended to investigate whether Machine Learning (ML) ensemble techniques can be used to improve on the performance exhibited by the existing systems beyond that of each individually. Ensemble techniques are generally defined as those that combine the results of several individual ML algorithms. 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 3.

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 off-the-shelf

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

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 5.1 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, is it 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 4, and compared with the underlying input systems in Figure 4.

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 4: Ensemble Approach Accuracy for Underage Subjects.

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

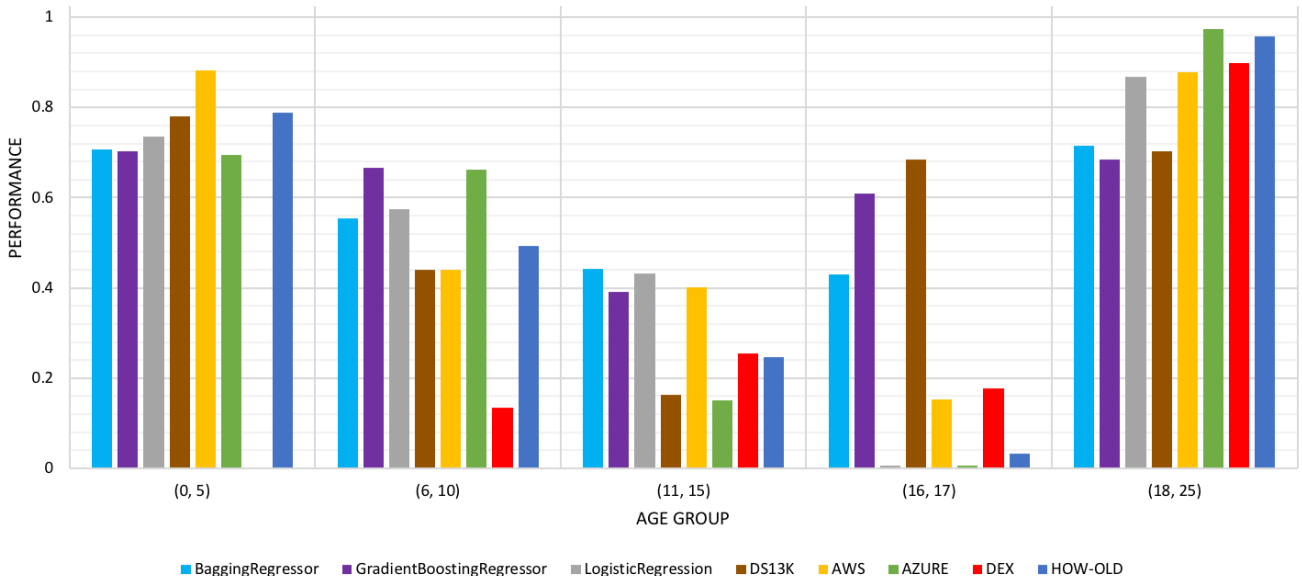

Figure 4: Performance vs Age Group.

range [18-25]. It can be observed in Figure 4 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 5.1, they are shown to generally overestimate a subject's age in this range, thus frequently misclassifying them as adults. Furthermore, the results in Section 5.1, specifically Figure 2 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, DEX's 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 youngster 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 [19], 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 CONCLUDING REMARKS

The four services evaluated in this study where Amazon Rekognition (AWS), Microsoft Azure, Deep Expectation (DEX), and How-Old.net. Initial evaluation results on the age range 0 to 25 years indicated that AWS had the overall lowest error rate, followed by How-Old.net; however, the ages that surround the borderline between minority and adulthood (considered to be 18 for this study) were found to follow a different pattern, where DEX surpassed the performance of AWS and Azure. Furthermore, an additional model named DS13K, based on VGG-16, was trained for this task. This achieved the highest accuracy for the borderline age range (16-17) when compared to the four other systems. Experiments on this

dataset indicated that ensemble approaches based on regression substantially outperformed the four systems used for this test, both in terms of mean absolute error and the task of classifying subjects into appropriate age ranges. Gradient Boosting and Bagging Regressor approaches outperformed the best individual system (DEX) for the key borderline range (16-17) by over 40%. This result offers a strong argument in favour of the proposition that ensemble learning has great potential in improving the precision of facial age determination.

Overall, even off-the-shelf regression techniques have been demonstrated to improve upon the performance of commercial offerings, by combining their outputs effectively. This offers a motivation for further work on bringing AI-based techniques to bear on this and other digital forensic challenges.

6.1 Future Work

Our aim is to investigate how to aid digital forensic cases with automated machine learning based techniques. Our objective is to expand this study further through comparative analysis of additional services. We have identified a need for higher-volume datasets for child face recognition to improve our models; 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.

REFERENCES

[1] Felix Anda, David Lillis, Aikaterini Kanta, Brett Becker, Elias Bou-Harb, Nhien An Le Khac, and Mark Scanlon. 2019. Improving the accuracy of automated facial age estimation to aid CSEM investigations. *Digital Investigation* 28 (2019), S142.

[2] Felix Anda, David Lillis, Nhien-An Le-Khac, and Mark Scanlon. 2018. Evaluating Automated Facial Age Estimation Techniques for Digital Forensics. In *12th International Workshop on Systematic Approaches to Digital Forensics Engineering (SADFE), IEEE Security & Privacy Workshops*. IEEE.

[3] Modesto Castrillón-Santana, José Javier Lorenzo Navarro, and Cristina Freire Obregón. 2016. Boys2Men, an age estimation dataset with applications to detect enfants in pornography content. (2016).

[4] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. 2017. Using Ranking-CNN for Age Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2009. EigenTransfer: A Unified Framework for Transfer Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. ACM, New York, NY, USA, 193–200. https://doi.org/10.1145/1553374.1553399

[6] Antitza Dantcheva, Carmelo Velardo, Angela D'Angelo, and Jean-Luc Dugelay. 2011. Bag of soft biometrics for person identification. *Multimedia Tools and Applications* 51, 2 (01 Jan 2011), 739–777. https://doi.org/10.1007/s11042-010-0635-7

[7] Yuan Dong, Yinan Liu, and Shiguo Lian. 2016. Automatic age estimation based on deep learning algorithm. *Neurocomputing* 187 (2016), 4–10.

[8] Eran Eidinger, Roee Enbar, and Tal Hassner. 2014. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security* 9, 12 (2014), 2170–2179.

[9] Jason Farina, Mark Scanlon, Nhien-An Le-Khac, and M-Tahar Kechadi. 2015. Overview of the Forensic Investigation of Cloud Services. In *10th International Conference on Availability, Reliability and Security (ARES 2015)*. IEEE, Toulouse, France, 556–565. https://doi.org/10.1109/ARES.2015.81

[10] Eilidh Ferguson and Caroline Wilkinson. 2017. Juvenile age estimation from facial images. *Science & Justice* 57, 1 (2017), 58–62.

[11] Andrew P Founds, Nick Orlans, Whiddon Genevieve, and Craig I Watson. 2011. Nist special databse 32-multiple encounter dataset ii (meds-ii). *NIST Interagency/Internal Report (NISTIR)-7807* (2011).

[12] Y. Fu, G. Guo, and T. S. Huang. 2010. Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 11 (Nov 2010), 1955–1976. https://doi.org/10.1109/TPAMI.2010.36

[13] Google. 2018. Using AI to help organizations detect and report child sexual abuse material online. https://www.blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online/

[14] Petra Grd and Miroslav Bača. 2016. Creating a face database for age estimation and classification. In *Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2016 39th International Convention on*. IEEE, 1371–1374.

[15] Hu Han, Charles Otto, and Anil K Jain. 2013. Age estimation from face images: Human vs. machine performance. In *2013 International Conference on Biometrics (ICB)*. IEEE, 1–8.

[16] Juliane A Kloess, Jessica Woodhams, Helen Whittle, Tim Grant, and Catherine E Hamilton-Giachritsis. 2017. The challenges of identifying and classifying child sexual abuse material. *Sexual Abuse* (2017), 1079063217724768.

[17] Quan Le, Oisín Boydell, Brian Mac Namee, and Mark Scanlon. 2018. Deep Learning at the Shallow End: Malware Classification for Non-Domain Experts. 26 (07 2018), S118 – S126. https://doi.org/10.1016/j.diin.2018.04.024

[18] Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 34–42.

[19] David Lillis, Brett Becker, Tadhg O'Sullivan, and Mark Scanlon. 2016. Current Challenges and Future Research Areas for Digital Forensic Investigation. In *The 11th ADFSL Conference on Digital Forensics, Security and Law (CDFSL 2016)*. ADFSL, Daytona Beach, FL, USA, 9–20.

[20] Khoa Luu, Keshav Seshadri, Marios Savvides, Tien D Bui, and Ching Y Suen. 2011. Contourlet appearance model for facial age estimation. In *Biometrics (ijcb), 2011 international joint conference on*. IEEE, 1–8.

[21] Sumit Mund. 2015. *Microsoft azure machine learning*. Packt Publishing Ltd.

[22] P Jonathon Phillips, Harry Wechsler, Jeffery Huang, and Patrick J Rauss. 1998. The FERET database and evaluation procedure for face-recognition algorithms. *Image and vision computing* 16, 5 (1998), 295–306.

[23] M Ratnayake, Z Obertová, M Dose, P Gabriel, HM Bröker, M Brauckmann, A Barkus, R Rizgeliene, J Tutkuviene, Stefanie Ritz-Timme, et al. 2014. 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 (2014), 803–808.

[24] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2016. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision (IJCV)* (July 2016).

[25] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2008. RUSBoost: Improving classification performance when training data is skewed. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 1–4.

[26] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[27] Frank Wallhoff. 2006. Facial Expressions and Emotions Database. http://www-prima.inrialpes.fr/FGnet/html/home.html

[28] Sun-Chong Wang. 2003. Artificial neural network. In *Interdisciplinary computing in java programming*. Springer, 81–100.

[29] Economy Watch. 2010. US Economy. *Economy Watch* (2010).

[30] Heidi Weber, António Cruz Rodrigues, and Américo Mateus. 2016. Emotion and Mood in Design Thinking. *Design Doctoral Conference'16: TRANSversality - Proceedings of the DDC 3rd Conference* July (2016), 65–72.

[31] Song Yang Zhang, Zhifei and Hairong Qi. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.