



Can Amazon Rekognition Be Used to Reliably Compare Driving Licence Photos for Identity Verification?

Salah Mohamed

Supervisor: Mark Huntly

Project submitted as part of the requirements for
the award of MSc Software Engineering Web
Development

January 2026

DECLARATION

I, Salah Mohamed declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.



Signed :

26/01/2026

Date :

FORM OF CONSENT

I Salah Mohamed hereby consent that my Project, submitted in candidature for the MSc Software Engineering Web Development degree, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.



Signed :

26/01/2026

Date :

Abstract

This study looks at whether AWS Rekognition can reliably verify identity by comparing driving licence photos, specifically for the DVLA's Ten-Year-Renewal process. Currently, DVLA caseworkers manually review thousands of photos daily, a time-consuming task prone to human error and causing delay on application outcome as it sometimes takes weeks to manually review photos depending on queue, particularly when comparing images taken a decade apart as people age.

To experiment facial recognition in DVLA context a serverless application is used for biometric matching and checking photo standards. The tests included 359 genuine pairs and 30 impostor pairs from AgeDB, MORPH, CACD, and FGNET datasets. Combining photos from these datasets allowed a good spread of demographics and age gaps, some stretching up to 23 years between photos.

The CompareFaces API performed well it averaged 99.92% similarity with 99% confidence and correctly rejected all impostor pair. Even with photos taken 10 years apart the exact scenario DVLA faces with licence renewals it still managed 99.80% average similarity, which shows it handles natural aging reasonably well. That said, the demographic breakdown wasn't quite as uniform. Mixed Race males in their 40s saw accuracy decline to 98.66%, while East Asian males aged 50-60 declined to 99.23% both the lowest scores recorded. Female subjects showed much more consistent results variance of just 0.08% compared to males 0.12% variance.

Photo standards validation revealed significant limitations. While the APIs performed well at detecting multiple faces, glasses, headwear, and pose violations, it performed poorly with smile detection with a 40% FAR rate, occlusion detection with 50% FAR rate, and eyes detection 12.5% FAR rate. With error rates this high, there's no way DVLA could rely on automated photo standards checking without caseworkers reviewing the standards manually.

Based on the findings, Rekognition is better suited as a support tool for caseworkers rather than deciding on its own. A hybrid setup makes more sense let the algorithm do what it's good at consistent pattern analysis while keeping caseworkers in the loop for tricky cases and those demographic groups where performance wasn't quite as stable. As for photo standards checking, that really needs to stay manual for now. The only alternative would be building custom models trained specifically on DVLA's requirements, but that requires lots of resources.

- Abstract
- Table of Content
- List of Figures
- List of Tables
- Acknowledgements

Table of Contents

CHAPTER 1 – INTRODUCTION.....	8
1.1 Research background	8
1.2 Research Statement.....	10
1.3 Research Aim	10
1.4 Research Objectives	10
CHAPTER 2 - LITERATURE REVIEW.....	11
2.1 Introduction	11
2.2 Historical Development of Facial Recognition Technology	11
2.2.1 Early Development 1960s-1980s	11
2.2.2 Breakthrough (1990s-2000s).....	12
2.2.3 Deep Learning Revolution (2010s)	12
CNN Architectures and Breakthroughs.....	12
2.2.4 Modern Commercial Applications	14
2.2.5 Modern Advances	14
2.3 Facial Recognition Methods and Process	16
2.3.1 Core processing Steps	16
2.4 Performance and Evaluation Methods.....	17
2.4.1 Standards Accuracy Measurements.....	17
2.4.2 Evaluation of Demographics Fairness	17
2.4.3 Procedure and Benchmarks Datasets.....	18
2.4.4 Performance Metrics.....	18
2.5 Challenges in Facial Recognition	19
2.5.1 Age Progression	19
2.5.2 Environmental Challenges	20
2.5.3 Biological Challenges	20
2.6 Algorithmic and Social Challenges	21
2.6.1 Bias and Fairness Issues	21
2.6.2 Datasets Bias Challenges.....	22
2.6.3 Synthetic Approaches	22
2.6.4 The UK Demographic	22
2.7 Legal and Ethical Considerations	24

2.8 Regulatory and Standards Frameworks	25
2.9 Cloud based Facial Recognition Services	27
2.9.1 Technical Implementation Frameworks and System Architecture	27
2.9.2 Performance Evaluation Under Real-World Conditions	28
2.9.3 Comparative Analysis of Cloud Provider Capabilities	29
2.9.4 Scalability, Cost-Effectiveness and Deployment Considerations	30
2.9.5 Privacy Considerations in Cloud-Based Facial Recognition Systems.....	31
2.10 Government Facial Recognition Deployment Challenges	32
2.10.1 International Legal Examples	32
2.11 Research Gaps	34
<i>CHAPTER 3 - RESEARCH METHODOLOGY</i>	<i>35</i>
3.1 Research Design	35
3.2 Application Architecture.....	36
3.2.1 Processing Functions	38
3.3 Software and Hardware Requirements.....	39
3.4 Research Data Analysis and Metrics Evaluation	40
3.4.1 Data Analysis	40
3.5 Ethical Considerations.....	41
3.6 Research Materials	42
3.6.1 Research Data.....	42
3.6.1.1 Sample Size and Quality Control	42
3.7 Chapter Summary	43
<i>CHAPTER 4 – EXPERIEMTS AND RESULTS.....</i>	<i>44</i>
4.1 Experiment Setup	44
4.1.1 Comparison Request Processing Flow.....	44
4.2 Results and Analysis.....	48
4.2.1 Analysis of Demographic Performance	48
4.2.2 Age Gap Impact Analysis	54
4.2.3 Photo Standards Compliance Analysis	57
4.2.4 Serverless Performance Analysis	59
4. 3 Discussion	60
4.4 Chapter Summary	63
<i>CHAPTER 5 – CONCLUSION, LIMITATION, AND RECOMMENDATIONS.....</i>	<i>64</i>
5.1 Summary	64
5.2 Conclusion	65
5.3 Limitation	66
<i>CHAPTER 6 – REFLECTIONS</i>	<i>67</i>

REFERENCES	68
-------------------------	-----------

List of Figures

Figure 1 CNN	13
Figure 2 Core Steps	16
Figure 3 Research Design.....	35
Figure 4 Application Structure	36
Figure 5 Verification Request Example	44
Figure 6 Verification Response Example	45
Figure 7 Photo Standards Request.....	46
Figure 8 Comparison data sample.....	56
Figure 9 Data Sample (Photo Standards).....	58
Figure 11 Gantt Chart Proposal	73
Figure 12 Gantt Chart	73

List of Tables

Table 1 Classification Function	41
Table 2 Standards Classification Function	41
Table 3 Overall Performance	48
Table 4 Security FAR	48
Table 5 Gender Comparison	49
Table 6 Ethnic Groups	49
Table 7 Female Demographics	50
Table 8 Male Demographics	51
Table 9 Age Analysis.....	52
Table 10 Gap Analysis.....	54
Table 11 Photo Standards	57
Table 12 Serverless Performance	59

Acknowledgement

I would like to thank my supervisor, Mark Huntley, for all the help and advice he gave me during this project. His feedback helped me develop my research approach and improve the overall quality of my work. Mark always pointed me in the right direction when I wasn't sure how to tackle something, and his guidance made a real difference in getting this dissertation finished.

CHAPTER 1 – INTRODUCTION

1.1 Research background

Everyday the internal caseworkers at the Driver and Vehicle Licencing Agency reviews and verify thousands of identify photos as part of driving licence applications for example new licences, applications for updating photo on licence and updating a mandatory photo as part of Ten-Year-Renewal driving licence applications. Photo verification is an important part of this process. Each submitted photo is reviewed manually by a caseworker the verification processes is time-consuming, require a lot of work, and there were lots of cases where caseworkers falsely accepted photos as true likeness and rejected photos that are valid , especially when comparing Ten Year Renewal application photos where photos are often taken ten years apart as people naturally age and their facial features change over time making this process more difficult for hum reviewers.

Experiments have demonstrated significant differences in facial recognition abilities among people. It classified people into specific categories of facial recognition abilities, ranging from super recognisers who have exceptional ability in identifying faces under various conditions to individuals who find basic face recognition challenging [1]. Super recognisers represent a tiny portion of the population and majority of people including caseworkers poses low to moderate facial recognition abilities that declines in challenging conditions such as age progression [2] [3].

In recent years the use of facial recognition has become a game changing way for automating the verification of identity. Facial recognitions have evolved over time going from a basic geometric feature matching to a more advanced deep learning architectures that can deal with complicated situations in real-world setting [4]. AWS Rekognition and other available cloud-based facial recognition solutions are good options for applications that need to process a lot of biometric data quickly and affordably. [5].

As DVLA currently employs serverless designs across its application integrating facial recognition software opens new ways to make identity verification application that is fast and can scale with the needs of the business. Serverless frameworks allow for seamless scaling, lower operating expenses, and pay-per-use rates model that fit within DVLA budget limits while still meeting high performance and reliability standards [5] There are issues when it comes to making sure that accuracy, fairness, and compliance are maintained across different

demographic groups and at different times. Age progression effects, demographic bias, and compliance with international photo standards are essential elements that require thorough assessment to ascertain the feasibility of automated facial recognition for DVLA applications.

1.2 Research Statement

The DVLA present way of verifying photos depend heavily on caseworkers checking photos manually for matches against existing photos stored in the system for Ten-Year-Renewal applications which slows down the processing of applications and causing delays while a comparison is in a queue for review. Particularly the Ten-Year-Renewal makes these problems worse because it becomes harder to verify customers identity when their appearance changes a lot as they get older. Cloud based facial recognition providers show lots of potential and may be used to automate this process that can improve applications processing time, but we still don't know how well they work for the DVLA specific needs. One of the gaps is in our understanding of how AWS Rekognition service works when it comes to things like age progression, fairness across demographics, and checking if photos meet international photo standards.

1.3 Research Aim

This experimental research aims to evaluate the effectiveness of AWS Rekognition for DVLA photo verification by developing and testing a serverless facial comparison application, and measuring its appropriateness, accuracy reliability, fairness, and compliance with DVLA photo standards and under various condition like age progression, fairness across different demographics groups.

1.4 Research Objectives

Obj1 – To critically analyse the current literature on the use facial recognition technology, identity verification, age progression, fairness, and ISO/ICAO photo standards to identify research gaps and future work.

Obj2 – To identify a suitable research methodology defining photo datasets, ethical consideration, experimental designs and evaluation metrics.

Obj3 – To design and develop a serverless facial comparison application using serverless framework, AWS Rekognition, API Gateway, Lambda, S3, DynamoDB.

Obj4 – To evaluate the performance of the developed application by measuring identify verification accuracy, fairness across various age groups, and diverse demographic groups.

Obj5 – To assess scalability of the developed application by analysing speed and usability in DVLA identity verification providing recommendations and improvements.

CHAPTER 2 - LITERATURE REVIEW

2.1 Introduction

Face recognition has gained popularity over the last 30 years because it simplifies analysis of images and pattern detection. At least two factors explain this such as trend in the wide range of commercial and legal needs, and the availability of relevant tools like smart phones, cameras. Even while machine learning solutions have advanced, the way they perform is limited to real-world settings. Identifying face images in an uncontrolled setting such lighting, posture, facial expression, occlusion, disguises etc. remains difficult [6].

Probably the face is the most common biometric feature. Since photography emerged, government agencies and commercial companies are keeping facial photos. As referential databases, these collections have matched and compared offender, eyewitness, and victim facial photos in criminal investigations.

Face recognition has an extensive history; it involves computer scientists and many more. The face of an individual is not a perfect method in relation to other biometric features; it generally demonstrates lower accuracy than methods like iris or fingerprint recognition and may be affected by cosmetics, disguises, and lighting conditions. Nevertheless, the face possesses advantages that position it as one of the most preferred biometric features for identity verification [6] [7].

The face is a realistic biometric feature used by people for recognition, making it the most used biometric features for verification and authentication [7]. Facial images may be taken swiftly without physical touch, making them more comfortable than fingerprint or iris images. Also, a facial recognition device may gather data in a nice, accepted way [7].

2.2 Historical Development of Facial Recognition Technology

2.2.1 Early Development 1960s-1980s

FRT is quite important for identity verification, and it started in the mid-sixties where US researchers team studied facial recognition computer programming. They came up with a semi-automatic approach, in which operators enter twenty computer-generated measurements, like the dimensions of the mouth or eyes. This approach was further enhanced in the late seventies by the addition of twenty-one extra indicators like lip width, hair colour etc. [4].

The late eighties saw the introduction of AI to programmers which was introduced to develop previously utilised theoretical tools, which exhibited numerous limitations. Mathematics specifically, linear algebra was employed to analyse images from alternative perspectives and to develop methods for their simplification and manipulation without reliance on human markers [4].

2.2.2 Breakthrough (1990s-2000s)

in the early nineties researchers from the Massachusetts Institute of Technology introduced the initial effective implementation of facial recognition technology, Eigenfaces, which employs the statistical Principal Component Analysis technique [4].

Towards the end of the decade more progress was made within industry and academia on this subject, the Defence Advanced Research Projects Agency established the Face Recognition Technology program, which supplied the global community with a substantial and challenging database consisting of 2,400 images representing 850 individuals [4].

In 2005: The Face Recognition Grand Challenge competition was initiated to promote the advancement of face recognition technology aimed at supporting existing facial recognition initiatives [4].

2.2.3 Deep Learning Revolution (2010s)

At the end of the decade saw progress speeds significantly with the arrival of deep learning, a machine learning approach founded on artificial neural networks [9]. The computer chooses the points to be compared and its learning is enhanced when extra images are supplied [4].

CNN Architectures and Breakthroughs

In 2012 Alex Net a deep CNN was introduced which won the ImageNet Large Scale Visual Recognition Challenge achieving higher accuracy than others. This marked the start of CNN revolution in computer vision. It showed that deep learning architectures could learn hierarchical feature representations on their own, without the need for manual feature engineering. This changed the way facial recognition works in a big way [8].

CNN Processing Pipeline

CNN analyse facial images using several layers which automatically extract features using numerical operations. Process is done layer-by-Layer starting with the initial input Layer which takes facial image with RGB colour channels and pixels. Then this is passed to a convolutional layer which use learnable filters and kernels to find things like edges, textures, and patterns by

running convolution operations. Activation Functions adds non-linearity, which enables identifying complex patterns. Pooling Layers keeps key characteristics while reducing the size of the space. Dropout Layers used during training, randomly turning off neurons stops overfitting. Fully Connected Layers Put together the features that were taken out to finish classifying or verifying a face [4].

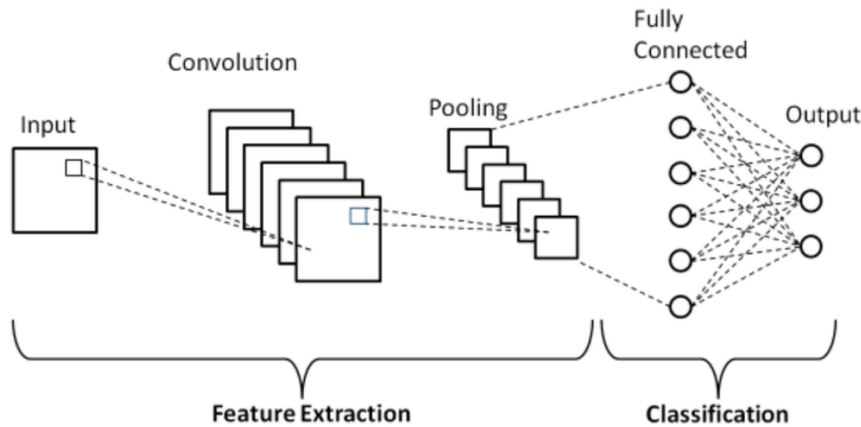


Figure 1CNN

VGG networks made facial recognition better by adding a few important new features such as ore complex feature learning is possible with deeper architecture, which has 16 to 19 layers. Smaller filters only about 3×3 convolutions capture more detailed facial features and adding more nonlinearity. Also, the network has the same 3×3 filter sizes and stride 1 everywhere and it has multiple convolutions using more than one 3×3 filter makes the receptive fields bigger like 5×5 or 7×7 . VGG demonstrated that network dept h was more important than filter size for facial feature extraction [10].

The introduction of ResNet in 2015 mitigated the vanishing gradient problem that made deeper networks training challenging. This made it possible to train networks with more than 100 layers [59]. ResNet's skip connections let gradients flow directly through the network. This made it possible to build much deeper architectures without losing performance. This breakthrough was especially important for facial recognition because deeper networks could learn more complicated representations of facial features. This made the accuracy of facial recognition systems much better on challenging datasets like LFW and IJB-A [59]. Many of the best face recognition systems are built on ResNet architectures.

2.2.4 Modern Commercial Applications

The advancement in the field saw the introduction of many commercial applications. Facebook's DeepFace, developed in 2014, achieved over 97.35% accuracy on the Labelled Faces in the Wild (LFW) dataset, which is almost as good as human performance [4].

The year 2015 saw further enhancement when Google's introduced its groundbreaking FaceNet, which changed facial recognition industry through directly from face images to a compact Euclidean space in which distances showed how similar the faces were [4]. Also, the approach performed higher compared to previous techniques achieving excellent scores on LFW and YouTube Faces DB, showing how powerful deep learning metric techniques can be [4].

ArcFace was created which implemented additive angular margin loss, that improved the ability of facial features to tell the difference between people by increasing boundaries of decision in the angular space. This method got the best results on several benchmarks and became the most popular choice for commercial facial recognition software because it was more accurate and reliable [4].

2.2.5 Modern Advances

2.2.5.1 Transformer-Based Architectures

Vision Transformers (ViTs) have improved the field by achieving excellent results when compared with traditional CNNs in challenging situations like pose variations. The Face Transformer architecture incorporated self-attention mechanisms which is effective in capturing long-range dependencies in facial features, improving performance on age progression face verification. Hybrid CNN-Transformer models integrate local feature extraction with global context comprehension, which is effective in the temporal variations related to cross age verification [11].

2.2.5.2 Bias Mitigation and Fairness

Recent Advancement made it more important to handle demographic bias in facial recognition systems. Calling for more dataset with balanced demographic representation in terms of age, gender, and ethnicity, which made model training fairer [12]. Also, Adversarial debiasing methods today effectively reduce differences in performance among demographic categories throughout training. Algorithmic auditing frameworks offer standard approaches for identifying and quantifying bias in operational systems, fulfilling the compliance mandates highlighted in facial recognition shortcomings [13].

2.2.5.3 Explainable AI in Face Recognition

For applications that require algorithmic transparency its essential to use explainable AI techniques. Grad-CAM visualization techniques show which parts of the face affect recognition decisions to know biased decision patterns. Heatmaps are used to give reasons for matches and unmatched, which helps meet the algorithmic accountability standards set by new AI governance frameworks [14]. These tools for explaining decision making can help in handling bias challenges and ensuring that different applications are transparent and follow guidelines [15].

2.2.5.4 Detection of Liveness and Anti-Spoofing

Advanced presentation attack detection techniques can stop advanced spoofing like deepfakes. Multi-modal uses RGB analysis, infrared thermal imaging, and depth sensing to tell if a real human is present. Physiological detection checks for liveness in real time by checking for micro movements patterns, and eye movements [16]. Recent advances in adversarial training make it less vulnerable to new attack methods, which helps keep security effective as spoofing techniques get better.

Today, today FRT reached major technical advances in neural network architectures, optimization algorithms, and deployment strategies which reached human-level performance allowing it to be used in both commercial and government applications. Advancement in areas like deep learning models, edge computing solutions for real-time processing, and anti-spoofing mechanisms. Allowing it to deploy in filed such as passport control, healthcare, law enforcement, retail etc. These advancements led to social issues like privacy, algorithms fairness, abiding laws, and balance between security and civil liberties. These introduced new policies and ethical AI frameworks.

2.3 Facial Recognition Methods and Process

2.3.1 Core processing Steps

Face detection and normalization, feature extraction, and classification are the three main steps in automated face recognition. Face detection finds faces in an image and tells if there are any. Feature extraction makes a signature vector that uniquely identifies each face by its unique features. Classification either verifies matches one face to another to control access or identifies compares a face to several stored faces. There can be some overlap between these steps. For example, facial features used for extraction are often used for detection at the same time [4].

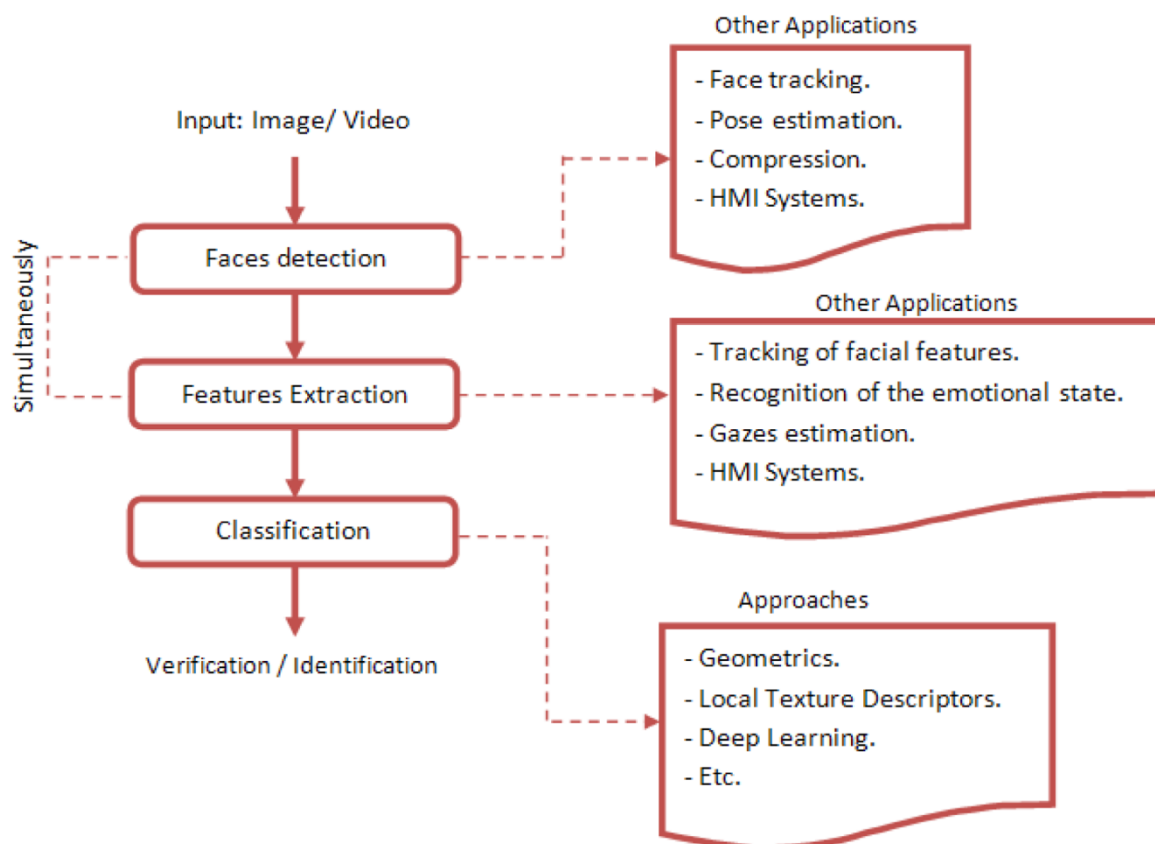


Figure 2 Core Steps

2.4 Performance and Evaluation Methods

2.4.1 Standards Accuracy Measurements

The examination of facial recognition technology demands standard measurements and procedures to provide accurate performance measurement under numerous conditions and demographic groupings. For government purposes, it is especially crucial that things are accurate, fair across different groups, and consistent, since these things have a big impact on government services and following the law.

Facial recognition system efficiency is typically tested using the False Acceptance Rate (FAR) and False Rejection Rate (FRR), that estimate the risk of incorrectly approving an impostor or disapproving a valid match, respectively [17] [18]. However, using these metrics may mask demographic disparities for example, Rekognition may achieve 1% overall FAR while having 10% FAR for a certain ethnic group which end up giving unfair results that are not revealed.

The Equal Error Rate (EER), which is the same as the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), is a single value that may be used to compare the performance of different solutions at different points of operation. To balance user experience and security, identity verification requires low FAR and FRR [17] [18].

Receiver Operating Characteristic graphs indicate the compromise among FAR and FRR at various decision thresholds, allowing system operators to determine the appropriate operation points depending on unique needs. The Area Under the Curve is used as a performance metric that doesn't rely on the threshold value. For critical applications uses, high values reflect excellent performance [17] [18].

In general, solutions need one-to-one verification, like comparing a saved image to a provided image, rather than one-to-many identification, like searching a database. The verification reliability is measured by computing True Accept Rate (TAR) at FAR levels [17] [18].

When systems give confidence scores or more than one candidate match, rank-based metrics become relevant. Rank-1 accuracy gives you how many times the proper match comes up as the top result, and Cumulative Match Characteristic (CMC) curves show how many times a match is identified at different levels [18].

2.4.2 Evaluation of Demographics Fairness

Demographic parity requires consistent system performance across diverse demographic groups, evaluated by comparing the False Acceptance Rate (FAR) and False Rejection Rate

(FRR) among gender, age, and ethnicity categories. The Demographic Differential indicator measures performance gap between demographic groups. [16].

Equalised odds evaluation makes sure that both the false acceptance rate (FAR) and the false rejection rate (FRR) are the same for all demographic groups. Individual fairness measures, on the other hand, make sure that people with similar backgrounds get the same treatment, no matter what their demographic traits are. Recent studies have developed benchmark datasets that precisely represent the demographics of the community under examination for these evaluations [17]. But there is no common ground on which fairness metric should be the most important for government identity verification, so decisions about how to use them are not clear.

2.4.3 Procedure and Benchmarks Datasets

Established assessment procedures ensure producing a more consistent and accurate results throughout various investigations. The various facial recognition test procedure points out to the best ways to test facial recognition solutions. These procedures lay out the steps for preparing images, matching them, and reporting results so that evaluations are always the same [20].

LFW is one of the key benchmark datasets for verifying faces without any restrictions. CFP is a key benchmark dataset for testing how well a system can recognise faces in different poses. AgeDB is a key benchmark dataset for testing how well a system can recognise faces at different ages [19].

These widely used datasets suffer from demographic bias LFW contains predominantly white male subjects, while AgeDB lacks sufficient representation of diverse ethnic groups and elderly individuals over 70 this pattern is often associated with freely available datasets. Furthermore, dataset demographics often don't match real-world populations MORPH is mostly Black. While the UK population is diverse this may require more than one dataset to achieve objectives set in this study.

2.4.4 Performance Metrics

In addition to accuracy measures, serverless facial recognition applications require testing things like latency and resource utilisation. mostly, requests and response time must be efficient. [5]. Also testing the system scalability under different load, from having one user to many imitating online traffic. Load testing mirror usage scenarios during operation hours [5].

2.5 Challenges in Facial Recognition

2.5.1 Age Progression

Face recognition solutions frequently correctly recognise people, but they have difficulty identifying children as they become older or photos of a person taking years apart. As a person become older, their faces changes in ways such as skin tone, weight, hair, and beard [21] [22]. Different ethnic groups age in different ways. For example, Asian populations tend to keep their facial structure longer, while African populations show more obvious changes in skin texture. However, most studies don't take these differences into account over long periods of time.

Generally facial recognition when comparing an aged face is lessened as the period between photographs gets longer [23] [24]. Craniofacial growth is the term for this change, which happens during the creation and development stages of the head, neck, and face. Current research on age progression relies a lot on synthetic aging datasets made by GANs. However, these artificial changes might not accurately show how people age in real life, especially the small changes that happen in different ethnic groups over long periods of time.

Identifying an individual's future or previous appearance is a challenge that needs to be resolved in many areas such as in forensic investigations but also in facial recognition, biometrics, origin detection, entertainment, social networks, and cosmetics. As a result, investigating the topic of facial age change is critical. This also known as face aging, age progression, or face age synthesis, in addition to the lack of labelled data, several reasons, such as changing living situations, interventions, stress, and bad habits, complicate the issue.

Age progression remains as one of the primary issues with facial recognition solutions, especially for critical applications where lots of time have passed [21] [23]. The aging process in humans impacts people facial characteristics through mechanisms such as changes in the texture of the skin, decreased facial volume, and changes in facial structure, which can considerably affect facial recognition solutions accuracy [21] [23] [24]. The DVLA's ten-year driving licence renewal process is a good example of this problem. The applicants must submit in new photo which are compared to photos taken ten years ago, making things very hard to handle for human caseworkers and any proposed automated solution.

Work in computation age progression has advanced beyond basic morphing methods to a more complex deep learning method. The use of Generative Adversarial Networks (GANs) in

age progression have demonstrated good results in to generating some realistic age-progressed faces, in 2023 GANs were used and produced good results on datasets serving as benchmarks [25] [26]. However, bias in these datasets may affect how it works for different groups.

Age progression has significant impact on accuracy of facial recognition especially among various demographics and ages groups. Various studies have demonstrated that facial recognition accuracy generally drops as the period across enrolment and verification grows, with notably significant impacts for people aged of fifty years old [21] [23] [24].

2.5.2 Environmental Challenges

Uncontrolled lighting conditions have a big effect on facial recognition accuracy, and declines performance when the lighting is poor. Things like shadows, backlighting can affect facial details. Different light sources with different colour change the way a face appears, which makes it difficult for algorithms trained on standard conditions. In other situations, like very bright or dark lighting condition, add noise and prove harder to extract features therefore affecting the performance [6] [27].

Head rotation and non-frontal poses significantly diminish recognition accuracy, with performance deteriorating beyond ± 15 degrees from the frontal view. When you look at someone's profile, you can't see important features on both sides of their face. When you look at someone from a very high angle, the perspective changes. Changes in pitch, yaw, and roll change the expected spatial relationships between facial landmarks. Accuracy declines for poses that go beyond ± 30 degrees [6] [27].

Images with low-resolution pixels make it very hard to extract features, and for images smaller than 64×64 pixels, the accuracy drops below 50%. Motion blur and compression artifacts make it hard for texture analysis algorithms to work. Mobile cameras and surveillance systems often don't work well because of hardware limitations and automatic settings [6] [27].

Masks, sunglasses, and hair that cover part of the face can greatly affect performance by hiding important parts of the face. COVID-19 made it clear that covering the mouth and nose makes it harder to see faces, with accuracy dropping significantly. Sunglasses block out features about the eye region, and hats and hair hide important geometric cues [6] [27].

2.5.3 Biological Challenges

Medical conditions such as stroke, or facial paralysis can make the face look uneven, which can confuse recognition algorithms. Long-term identity verification can be hard because of

progressive conditions like changes that happen with age, weight changes, or treatments and cosmetics that change the structure of the face [6] [27].

Identical twins pose significant challenges of similarity, as their shared genetic traits result in nearly indistinguishable facial structures. Even fraternal twins and siblings who are very close can look enough alike to make false positive matches. Advanced systems need very specific features to tell genetically similar people apart [28].

2.6 Algorithmic and Social Challenges

2.6.1 Bias and Fairness Issues

The existing problems of biases in algorithms in the facial recognition solutions has attracted a lot more attention previously due to the popular research which showed that it performs differently across demographics groups. The groundbreaking research conducted has uncovered major differences in accuracy among demographic groups, showing error rates that differed by as much as a third across various gender and skin tone combinations [29]. Following this research an enormous amount of interest in the testing bias within facial recognition solutions which paved the way to laws being passed around the globe [30].

Recently completed thorough assessments have validated ongoing bias concerns among various facial recognition suppliers. In 2021 the US based agency NIST report on facial recognition Vendor Test showed big differences in false positive rates between different groups of people. This has impacts on solutions and law enforcement [20]. These results have prompted heightened examination of facial recognition implementation in critical solutions and demand for compulsory bias thorough assessment [30] [31].

There are many different reasons why facial recognition solutions have been biased. These include things like the use of unbalanced training data, selections made in the development of the algorithms, and the ways they are tested [30] [31]. It is widely established that various training datasets have previously excessively lacked comprehensive representations of diverse demographic groups, which resulted in models that produced poor accuracy on populations that are underrepresented. More recent research has concentrated more on designing bias mitigation strategies, such as adversarial debiasing, impartial training goals, and synthetic data augmentation [31].

The evaluation of fairness in facial recognition solution have advanced from basic accuracy analyses to a more sophisticated measurements of algorithmic fairness. Demographic equality,

equal odds, as well as individual fairness have become important ways to judge how fair a solution is [32]. Nonetheless, in research shows the process is quite complex to get fairness using several metrics concurrently, as it may require compromises among various fairness targets.

To make facial recognition solutions fair across different demographics, the two technical and social considerations need to be considered. Recent research showed the need of broad dataset curation, diverse teams of developers, thorough testing against all fairness criteria and continuous oversight of implemented solution [33]. Regulatory frameworks like the EU AI Act are starting to require fairness tests for AI applications that are high-risk, such as systems that use biometric data [34].

2.6.2 Datasets Bias Challenges

The EU suggested to put tough measures in place such as Artificial Intelligence Act along with comparable laws around the world have made it even more important to deal with bias and make sure that facial recognition systems are fair [34]. Recent studies have concentrated on formulating bias mitigation strategies and fairness-oriented training methodologies to tackle these issues.

2.6.3 Synthetic Approaches

Using generative models to make generate facial data has recently become an intriguing means to solve problems with datasets that are too small or biased. In 2021 work was carried with StyleGAN3 and showed that it can make photorealistic facial images that can be used to improve data sets for training purposes. Diffusion algorithms have recently demonstrated even more potential in making excellent synthetic faces that are more diverse and easier to control [25] [26].

Artificial data can be used for more than just simple data augmentation. The recent studies in this area showed that well-designed artificial datasets might enhance the accuracy of models and fix bias problems within various demographic categories. Using this technique can help in making a dataset that is well balanced datasets [25] [26].

2.6.4 The UK Demographic

When examining facial recognition applications for government use, it's important to understand what the UK population is like. The 2021 Census shows that UK is now more diverse. In 2001, white British people made up 87.5% of the population, but now they only

make up 74.4%. Asian groups make up 9.3% of the population, Black groups make up 4.0% of the population. People of mixed race make up 2.9% of the total [58].

In the past 20 years, the number of people from different ethnic background has more than tripled, from about 220,000 to over 920,000. Due to change in demographics, facial recognition applications that have been developed for populations not as diverse may have challenges with the UK's present scenario [58].

This may make it challenging for DVLA renewal process as commercial facial recognition solutions may not be trained on datasets that include the diverse UK's ethnic groups. The algorithms may not have seen enough examples of mixed-race people or people from smaller ethnic groups, which could lead to potentially higher error rates.

2.7 Legal and Ethical Considerations

As biometric data is sensitive and may be used for mass surveillance and algorithmic discrimination, it's important to consider ethical frameworks and that privacy protection [35].

GDPR states that facial recognition information constitutes special category sensitive information. This means that it needs explicit consent and additional safeguards. The Data Protection Act states that biometric data is critical and there are tough rules in place that governs how it can be processed, stored, and kept [35]. These laws require organisations and government agencies to adhere to confidentiality-by-design rules and carry rigours Data Protection Impact Assessments prior to putting facial recognition systems into use [35].

The black box design of today's deep learning powered facial recognition solutions makes it hard for solutions to be accountable and open about how it makes decisions [36]. The EU AI Act sets call for an explainable AI in developed solutions. This means that applications have a duty to explicitly clarify how algorithmic choices are reached [34] [37].

Researchers have been working on making facial recognition models that can explain their decisions while still being accurate. Work have carried to make facial recognition systems easier to understand using these methods like attention visualization, gradient-based explanations, and counterfactual analysis [15].

Privacy-by-design principles state that biometric systems should only collect and use the least amount of data needed for their intended purpose [38]. This means that for DVLA photo verification, data collection should be limited to facial comparison.

Recent studies have demonstrated that advanced and modern facial recognition solutions often retrieve excessive information beyond what is required for identity verification, encompassing age, gender, ethnicity, and emotional state [39]. The principle of purpose limitation requires that any extra data extracted must be explicitly justified and with proper consent acquired.

2.8 Regulatory and Standards Frameworks

2.8.1 International photo Standards

Adhering to the global guidelines set for identity document photos are essential to ensure that identification around the world can work together and are safe. The ICAO defines technical standards for identity documents. These standards include comprehensive standards for photos of faces used in biometric verification [39]. These standards set requirements for things like the photo resolution, lighting conditions, pose, and background colour that makes automated facial recognition work better [39].

The International Civil Aviation Organisation creates requirements thorough groups made up of aviation authorities, border control agencies, and biometric technology experts. These groups carry essential performance testing, the need for interoperability between countries, and security assessments [40] [41].

The ISO adds to ICAO standards with formats and quality standards for face image data exchange [44]. Groups like biometric vendors, agencies, and researchers create these standards. They provide detailed instructions on how to check the quality of images, such as how to measure sharpness, check exposure, and make sure the image is geometrically correct [42] [43].

The UK government's photo guidelines are examples of putting international standards into practice. The guidelines states facing forward, having neutral expression with the mouth closed, and the eyes must be clearly visible without any occlusion [60]. Technical specifications say that the head height must be between 29 and 34 mm within a 45x35 mm area, the backgrounds must be plain cream or light grey, and the printing must be of high quality on photographic paper. The rules cover common compliance problems like not allowing shadows, red-eye, facial coverings, and more than one person in a photo [60] [61].

There are special rules for kids under five no neutral expression needed and babies under one eye may be closed, but all other technical standards must still be met. These detailed requirements show how international biometric standards turn into specific instructions for how governments should check people's identities [44].

The recent developments and combining different APIs or models to make comprehensive photo standards checks to ensure that photos meet standards on a large scale. As these are already incorporated with AWS Rekognition like checking for eye direction, pose, and

background colour. Testing it to classify photos based on compliance with defined standards with ICAO may be possible.

2.9 Cloud based Facial Recognition Services

The introduction of online facial recognition completely changed how businesses use verification of identity, and surveillance. Amazon Rekognition is a powerful cloud-based service that offers robust face recognition capabilities [45]. It is arguably of the most popular solutions on the current market. It uses advance machine learning algorithms providing lots of facial analysis features far more than just matching faces. It can also extract emotion, gender, objects, and facial features [46].

Microsoft Azure Face API is an important player in this facial recognition. It uses advanced AI algorithms to discover, recognise, and analyse faces in images and provide various features such as and can be used for things like identity verification, touchless access control, and face blurring for privacy [47]. Azure's strategy is to use its full cloud infrastructure to offer facial recognition services that are suitable for businesses [48].

Google Cloud Vision is the third well known big cloud provider, but it takes a more focused approach. Its API allows its users to leverage machine learning techniques for image processing, such as image categorisation, face identification, text extraction, and many more. Similarly, the service focuses more on integrating with Google's larger machine learning ecosystem and has features for multiple objects detection and classification, as well as the location of each object inside the image [47].

There are lots of other services providers on the current market in addition to these giant ones. The five major commercial facial recognition solutions include Amazon Rekognition, Face++, Microsoft Azure, to show how many different options are out there. Each provider has its own strengths and focuses on different parts of facial analysis, from basic detection to more advanced emotion recognition [46].

The market has also seen mergers and changes in the availability of services. Some providers have stopped offering services or been bought by bigger companies. This is because the facial recognition industry is very competitive and the rules are always changing [46].

2.9.1 Technical Implementation Frameworks and System Architecture

Most modern cloud-based facial recognition solutions have complex architectures that often mix several services to handle a lot of users. Using AWS for facial recognition require integration of various services, such as Rekognition, Lambda, S3, API Gateway, DynamoDB

[45]. This distributed technology lets businesses build full facial recognition systems without spending a lot of money on infrastructure.

The serverless framework is appealing for facial recognition apps. The business logics are handled by Lambda functions. Lambdas are triggered by events, such as uploading images to S3 bucket, and handle tasks like preprocessing images, interaction with Rekognition, and database operations [45]. This event-driven architecture makes sure that resources are used in the best way possible and at the lowest cost.

AWS Recognition engine provides the main face analysis abilities that performs tasks such as face detection, comparison, and analysis. Rekognition is used during both the registration and authentication phases to ensure accurate identification and verification. Integrating DynamoDB acts as the database for storing user metadata, including face embeddings, user IDs, and other relevant information [45].

API management makes it easier to connect and use different systems. Amazon API Gateway facilitates communication between the client and the backend services. The API Gateway exposes RESTful endpoints that allow clients to interact with the face recognition system, such as submitting images for authentication [45].

2.9.2 Performance Evaluation Under Real-World Conditions

Evaluating facial recognition real challenging conditions revealed that its performance dropped when compared with controlled lab settings. Most research into facial emotion recognition has used high-resolution, front-oriented, full-face images. However, when images are collected in naturalistic settings e.g., using smartphone's frontal camera, these images are likely to be far from ideal due to camera positioning, lighting conditions, and camera shake [49]

Extensive testing has revealed numerous factors that influence the solution performance. Extensive performance testing was carried using five potential reasons for poor-quality natural photographs: rotation, partial face, brightness, blur, and noise. When not properly handled, each of these factors can make recognition accuracy much worse [49].

Performance metrics show how well cloud-based systems work when everything is perfect. The test showed that the recognition accuracy of 93.5% indicates that the system is highly effective in identifying faces from the test dataset [45]. But the accuracy can change a lot depending on the quality of the image and the environment conditions.

The speed of processing is an important measure of performance for real-world applications. The system demonstrated quick processing times, with an average of 320 ms per image under low load conditions. Scalability testing showed that systems can handle up to 100 images per second with near-maximum resource utilisation CPU 85%, Memory: 80% [45].

Different providers have very different effects on image distortions. Study discovered that the systems varied in how well they handled manipulated images that simulate realistic image distortion. Amazon, Baidu, and Face++ performed well and were stable to different angles. Their overall accuracies for the rotated images were almost as good as for the original images. However, Microsoft Azure and Affectiva were vulnerable to rotation manipulation [49].

However, the field is changing and recent studies shows that Rekognitio excel in comprehensive emotion detection capabilities, whereas Google's Cloud Vision provides a more restricted emotional range, potentially limiting its effectiveness in emotional analyses [46]. When it comes to biases, it's important to know the reliability of those providers suggests that facial recognition commercial APIs can show racial bias, with systems working differently for different demographic groups [50].

2.9.3 Comparative Analysis of Cloud Provider Capabilities

A thorough comparison shows that each major cloud provider has its own strengths and weakness. Thorough testing has shown that services work very differently from each other. AWS Rekognition and Azure Face API provide the detection of face as well as comparison whereas Google Cloud Vision provides face detection but not face comparison [47].

When comparing providers performance across several parameters, it's clear that they are different. Tests showed that AWS Face Rekognition performs better than others. In terms of parameters Amazon Rekognition proved itself remarkably well with the accuracy of 98.68% in poses, 97.645% in different lighting, 99.99% in expression, 99.68 in accessories and 98.76% in multi face recognition [47]. Azure Face API was 72.72% in poses, 86.69% in different lighting, 97.67% in expression, 95.42% in accessories and 83.72% in multi face recognition [47].

The ability to detect multiple faces shows big differences between providers. Findings revealed that when evaluating the tracking of a specific individual in a group of people, AWS Rekognition was able to recognise with a high confidence score whereas Microsoft Azure failed to recognise for most of the dataset samples [47]. This discovery has significant ramifications for surveillance and security applications.

The overall performance of the system shows that there are clear market leaders. The average confidence score of AWS Face Rekognition came out to be 98.95% which was higher than the Azure Face API with the accuracy of 87.24% [47]. Findings revealed that Rekognition-based framework when compared with other services offers higher accuracy rate, scalability and ease of integration [47].

Different providers have very different abilities to recognise emotions. Baidu scored highest overall accuracy value 95.8%, with Amazon second 86.8%, followed by Microsoft Azure 84.7%, Face++ 81.5%, and finally Affectiva [49]. The study also found that recognition systems rarely confused surprise, disgust, happy, and neutral expressions with other emotions, but fear and angry expressions were the hardest to identify and classify [49].

2.9.4 Scalability, Cost-Effectiveness and Deployment Considerations

Things to think about when it comes to scalability, cost-effectiveness, and deployment the economic benefits of cloud-based facial recognition systems have been a major reason why many businesses have started using them. The system is very scalable, has low latency, and is cost-effective because it uses cloud-based technologies and serverless architectures. This makes it perfect for real-time, large-scale deployments [45].

Because facial recognition apps use the serverless computing model in different ways, it has some benefits for them. The serverless architecture lets the system handle more users and images without having to make major changes or move resources around [45]. This flexibility means that businesses can handle different workloads without having to buy too many expensive computers.

Cloud deployment greatly lowers the need for infrastructure. Organisations can implement face recognition solutions without the need for extensive hardware or infrastructure investment [45]. because of the how the framework it made the technology available to smaller businesses.

Testing demonstrated the capacity to sustain high throughput with only minor increases in resource utilisation, indicating efficient resource management [45]. This proved that systems ability to handle up to 100 images per second with near-maximum resource utilisation while still performing well [45].

Different providers have different cost structures and adopting the pay-per-use model makes facial recognition available for both smaller and large production deployments as it only

charges per use. This pricing model has helped cloud-based facial recognition services become popular in many different fields and uses.

2.9.5 Privacy Considerations in Cloud-Based Facial Recognition Systems

Using cloud based FRT presents privacy issues that need to be considered carefully. Recent research demonstrates that Privacy concerns become even more pronounced when the FRT system is deployed in the cloud, where sensitive data is processed and stored by third-party providers [51].

Cloud-based facial recognition systems work in a way that everyone is responsible for them. The research makes it clear that the three cloud providers Microsoft Azure, AWS, and Google use the shared responsibility model for cloud security. This means that customers still own the data, but the cloud provider is responsible for managing the infrastructure and services that support the data [51].

Businesses using verification systems are still responsible for maintaining and securing their data, access control and identity management, configuring security settings for their applications and services, ensuring that their use of services is in accordance with relevant legal frameworks, like GDPR [51].

2.10 Government Facial Recognition Deployment Challenges

HMPO automated photo checking system was biased, rejecting photos of ethnic minorities and people with darker skin colour. Investigations discovered that the automated system was using stricter rules than human reviewers and that staff had to manually override legitimate photos that the system had rejected. It was systematically biased and clerks had to constantly approve rejected photos [52].

The court found South Wales Police's use of automatic facial recognition was against the law because it violated people's rights and didn't have enough protections against discrimination. The decision outlined that the technology was not tested properly for bias and that there was not a strong enough legal basis for its use. According to reports, the Metropolitan Police have a high false positive rate, and it disproportionately affects Black and ethnic minority communities [53].

The Home Office's visa application system demonstrated similar patterns of algorithmic bias to the passport office system, with higher rejection rates for applicants from certain ethnic backgrounds [54].

2.10.1 International Legal Examples

Facebook's automatic photo tagging feature used facial recognition to scan and identify users in uploaded photos without explicit biometric consent, violating Illinois' Biometric Information Privacy Act (BIPA). The lawsuit outlined that Facebook extracted and stored biometric information from users between 2011 and 2021 without their permission. Facebook ended up paying \$650 million in 2021. Facebook turned off facial recognition feature, which shows how big the legal and financial risks are of using biometric systems without proper consent frameworks [55].

Google Photos used facial recognition to help users with organising their photos without their consent. The lawsuit outline that Google used biometric data users' facial geometry to make face templates for photo organisation features. Google paid \$100 million in settlement fees. Not obtaining proper consent would lead to lawsuits and heavy fines [56].

Big Brother Watch took the UK government to the European Court of Human Rights over the police use of live facial recognition technology, as it broke privacy rights and the European Convention on Human Rights. The case showed how UK police used facial recognition to

watch mass people without a clear legal framework, scanning everyone in public places without any reason to believe they were doing anything wrong [57].

These challenges indicate a thorough bias testing is essential, strong legal frameworks, and obtaining proper consent before deployment. The fines and failures that both government agencies and private businesses faced shows how important it is to test the application like AWS Rekognition to make sure they follow the law and don't lead to unfair results.

2.11 Research Gaps

Looking at the current state of facial recognition research, several important gaps become clear when thinking about how these systems might work for government applications like DVLA's licence renewals.

Most studies on how faces change over time don't really tackle the specific challenge DVLA faces. There's age progression research for different facial recognition models, but it's mostly shorter periods or using a single benchmark dataset. Other studies tested Rekognition covered shorter periods and mostly object and face detections. But not covered gaps of 10 years that UK driving licences require and analysis and lacked analysis of age, ethnic groups. So, gaps include looking at longer period across different demographic.

The issues highlighted in HMPO system where it falsely rejected photos from ethnic minorities, so gaps focus on testing AWS Rekognition performance across different ethnicities, ages progression, and backgrounds found in diverse UK population.

The research that does exist on cloud facial recognition services tends to focus on general accuracy numbers without considering what government agencies need. There are gaps in testing if photos meet the strict photo standards that DVLA requires. This experiment aims to fill these gaps by testing AWS Rekognition for DVLA photo standards and identity verification.

CHAPTER 3 - RESEARCH METHODOLOGY

This section of the report outlines the research methods used in this experimental study to achieve the aims and objectives of analysing AWS Rekognition effectiveness for DVLA identity verification. It describes the serverless application, the proposed evaluation approach, performance evaluation metrics, research data and materials used for this experiment.

3.1 Research Design



Figure 3 Research Design

Figure 3 describes that this experiment uses the positivism philosophy and takes a deductive approach. This approach was chosen as this study is mainly quantitative and highly relies on gathering data, testing it and running a statistical analysis, and takes an objective evaluation of facial recognition accuracy, fairness, and compliance across different experiments and set of conditions.

The deductive approach was chosen because the experiment research tests existing facial recognition technology theories in identity verification for a driving licence photo context. This approach is beneficial as testing are planned ahead, and required data are gathered using annotated facial images datasets to assess AWS Rekognition service, establishing the effectiveness of the cloud service for DVLA Ten Year Renewal identity verification applications.

The research strategy chosen for this research is mainly based on experiment. To assess the use of AWS Rekognition's services accuracy and resilience for practical government use. facial

images datasets will be used to test Rekognition and assess its performance under different conditions such as age progression, demographic variations, and photo quality and photo standards checks.

This experimental research utilises a combination approach of different methods, such as qualitative and quantitative methods widely used within the present discipline of facial recognition technology.

The experimental research employs a cross-sectional time horizon to test the present AWS Rekognition service.

3.2 Application Architecture

In this section will highlight the proposed serverless application architecture that will mainly use various AWS services to achieve the aim and set of objectives. It also highlights how the key AWS services used are set up so that face recognition may be used in the process of checking, validating analysing and comparing DVLA driving licence images. The below figure 5 outlines the proposed application structure.

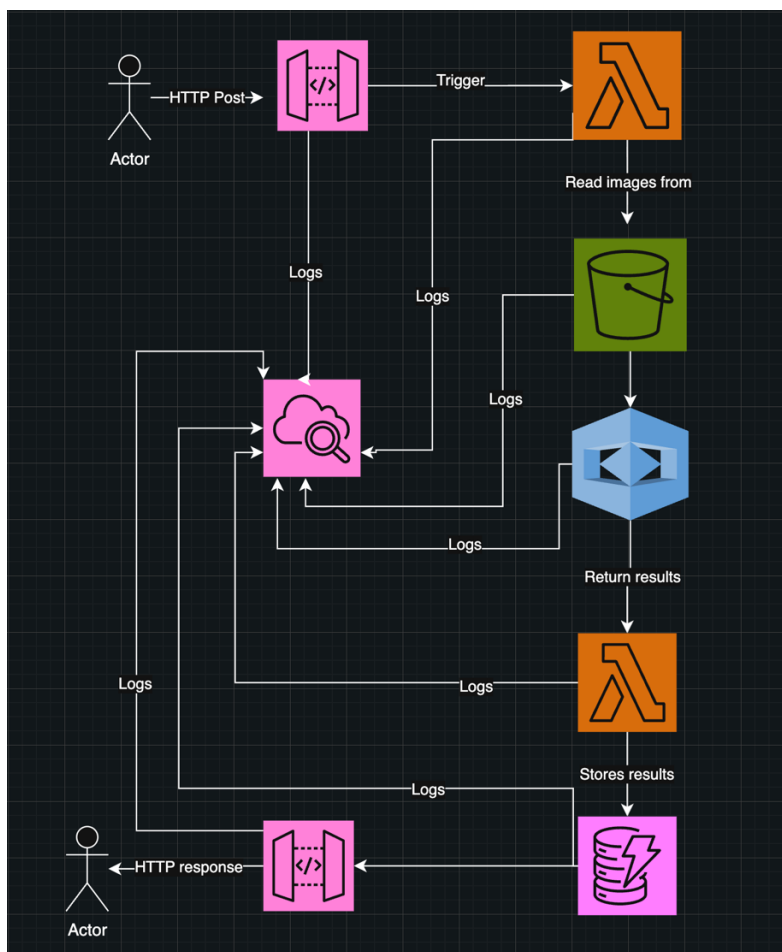


Figure 4 Application Structure

The Amazon API Gateway is a RESTful endpoint for managing requests and used as the application entry point. It handles request authentication, rate limits, as well as routing. It handles http requests that have image metadata in them and starts processing downstream using an event-driven architecture.

Face Comparison Lambda: This lambda checks for match between a source image and a target image and using a classification function that determines whether there is a match and based on similarity score it classifies it as approved for high similarity scores and if low score refers it to a supervisor for additional safety checks unapproved and refers it fraud investigations if low score or no match found.

Photo Standards Lambda: This lambda checks every submitted image against the set of requirements for a driver's licence by looking at things like the check for multiple faces in the image, the pose, wearing accessories, headwear, mouth open, face occlusion etc.

AWS S3: Two separate general buckets are used to safely store images used for comparison and checking image standards which Rekognition will use to run its analysis.

AWS Rekognition: Is the essential part that provides computer vision functions through a few API endpoints such as CompareFaces API which primarily checks for biometric matching with the similarity threshold set to 95% to verify identity with a high level of confidence. Another API used is the DetectFaces which primary looks at facial features like pose, eye and mouth state, accessory, and extracting demographic data such age and gender. Lastly another useful API used is the DetectLabels which looks at things that don't follow the rules, like headwear and background elements.

AWS DynamoDB: Is a NoSQL database that is used to create tables the first table is used for the image standards checks to keeps track of photo verification results, compliance scores, and details about violations and later be used for analysis. The second table is used to keeps track of demographic data, face comparison results, similarity scores, and classification results to be used for analysis later stages.

AWS CloudWatch: Is a useful service used to logs and monitoring on all the existing Lambdas, API Gateway endpoints, Rekognition API calls, and DynamoDB database operations helping in debugging and finding any issues.

3.2.1 Processing Functions

The classify images comparison function employs a three-tier risk assessment model for biometric matching decisions. High confidence matches scores with more than 95% similarity get automatic approval to speed up processing. Medium level matches between 50 - 94% similarity score are sent to a supervisor for review to find a balance between security and operational efficiency. No matches or low confidence similarity score start fraud investigation protocols.

The verify image standard compliance function runs a full validation to check all the defined compliance criteria using conditional checks. initially it ensures that only one face is detected on the image. Then it systematically checks facial features like mouth state, eye openness, sunglasses presence, pose, face occlusion levels, and headwear detection through object classification. Then every non-compliant detected is logged with specific reasons, and the function returns a binary approved or unapproved classification based on the return from the reasons.

3.3 Software and Hardware Requirements

This application mainly uses the Serverless Framework to set up and manage its infrastructure, and it runs on AWS cloud infrastructure using the Eu-west-2 region in London to ensure that data is kept within the UK to comply with the Data Protection requirements. Below are the hardware and software needed to complete this project.

HARDWARE

- Apple Mac pro laptop.

SOFTWARE

AWS Cloud Services:

- AWS Region: Eu-West-2 London for data residency compliance
- AWS Lambda: Node.js runtime with enough memory allocation
- AWS DynamoDB: With encryption enables and on demand payment mode
- AWS S3 Bucket: Storage
- AWS Rekognition: Comparison and image standards checks APIs
- AWS API Gateway: RESTful interface with authentication
- Testing APIs: Bruno to send http requests

DEVELOPMENT

- Programming language: TypeScript for type safety checks
- Framework: The Serverless Framework V3 for infrastructure as code
- Testing: Jest for unit tests
- Monitoring: AWS CloudWatch for debugging monitoring and logs

SECURITY AND COMPLIANCE

- Data Protection: GDPR
- Authentication: Setup policies and IAM roles
- Encryption: Use of AWS available encryptions

3.4 Research Data Analysis and Metrics Evaluation

This section covers how to analyse the experiment results for comparing driving licence photos and checking photo standards, plus the metrics I used to evaluate everything.

3.4.1 Data Analysis

Similarity measures the percentage match between two photos; higher scores indicate match. The API response typically returns similarity values from 0 to 100%. This was set to 95% threshold in this experiment which is a value used and recommended for critical applications to balance security and usability.

The APIs response include a Confidence level 0-100% indicating how certain in its similarity decision with values over 95% considered excellent for identity verification.

CompareFaces Evaluation Metrics

The evaluation includes standard biometric performance metrics:

True Accept Rate (TAR) = $\text{Correct Matches} / \text{Total Genuine Pairs} \times 100$

This is used to measures how the CompareFaces API correctly approves matching pairs.

False Accept Rate (FAR) = $\text{Wrong Matches} / \text{Total Impostor Pairs} \times 100$

This is used to measure security risk by calculating incorrect acceptance of different people as matches.

False Reject Rate (FRR) = $\text{Wrong Rejections} / \text{Total Genuine Pairs} \times 100$

This is used to measures user experience impact by tracking genuine pairs incorrectly rejected.

Photo Standards Evaluation Metrics

Classification Accuracy = $\text{Correct Classifications} / \text{Total Photos} \times 100$

False Acceptance Rate (FAR) = $\text{Violations Missed} / \text{Total Violation Images} \times 100$

This is critical for security as it measures how often non-compliant photos pass automated checks.

Classification Function

Similarity Score	Action	Description
$\geq 95\%$	Automatic Approval	High confidence match approved
50-94%	Supervisor Review	Manual verification required
$< 50\%$	Fraud Investigation	Low or no match potential security risk

Table 1 Classification Function

Photo Standards Classification

Classification	Criteria	Description
Approved	Meets standards requirements	Photo complies with all government standards
Unapproved	Violates one or more standards	Photo fails to meet required specifications

Table 2 Standards Classification Function

3.5 Ethical Considerations

This study adheres to ethical guidelines for handling biometric information and all photos used in this study came from publicly available datasets like AgeDB, MORPH, CACD, and FGNET which already have the right usage permissions.

The study adheres to GDPR rules by only using the minimum data needed just facial comparison features for testing identity verification. No other information was used beyond what's required for the analysis in this study.

Since these datasets are used widely in academic studies, there weren't any additional consent issues to worry about. The focus was purely on testing the technical performance of AWS Rekognition rather than collecting new biometric data.

3.6 Research Materials

This section study relies on 5 well-known facial recognition benchmark datasets, and these carefully chosen for the DVLA task.

3.6.1 Research Data

359 genuine comparison pairs and 30 impostor pairs from various datasets like AgeDB, MORPH, CACD, and FGNET. AgeDB was not enough on its own due to major bias requiring use of other datasets to ensure demographic coverage and the kind of age progression data required to test DVLA's age progression scenario.

The genuine pairs were filtered to include age gaps from 5 to 23 years, focusing more on that 10-year and above gaps that matches UK driving licence renewals. This allows investigating how aging affects Rekognition accuracy over the exact timeframe DVLA deals with. The subjects in the dataset range from 16 to 102 years old, covering most adult ages.

Balance across gender was in mind 179 female, 180 male pairs, ethnicity White, Black, Asian, East Asian, South Asian, Mixed Race, Hispanic/Latino, and Middle Eastern to represent most UK populations, and different age groups so I could properly test for fairness. However, the sample sizes were not equal due to bias in free datasets. White groups highest representation with 168 pairs but other ethnic groups had less data to fully assess potential bias.

For checking photo standards FRL dataset was used with a total of 87 photos across seven types of violations like smiling, mouths open 25 photos, multiple faces 4 photos, glasses 9 photos, face occlusion 8 photos, headwear 14 photos, closed eyes 8 photos, and wrong poses 12 photos. These represent the common ways people fall short in real licence applications, so allows testing how well DetectFaces and DetectLabels APIs handle DVLA photo requirements.

3.6.1.1 Sample Size and Quality Control

The 359 genuine pairs were chosen to give enough statistical power for spotting performance differences between demographic groups. This follows typical protocols for facial recognition evaluation that other researchers use.

Photo quality was at the centre any photos that were blurry, too low resolution, or had bad lighting got filtered out. Same thing for photos where faces were mostly covered or at weird angles that would make fair comparison impossible.

The uneven demographic representation wasn't ideal, but that's just how these free datasets are built. White populations ended up with the biggest sample while some ethnic groups had much smaller numbers. This is a known limitation that I had to work with, though it could be enough to spot major bias issues if they exist.

3.7 Chapter Summary

This chapter highlighted methods for testing how Rekognition works for DVLA photo verification. The experimental approach mixes quantitative biometric testing with demographic fairness analysis using datasets that other researchers have used before.

The application setup allows for scalable tests while keeping data within UK borders for compliance. Performance is measured using TAR, FAR, and FRR and demographic analysis to check if the APIs work fairly across different groups.

The methodology balances technical rigour and constraints working with existing datasets rather than collecting new data, using cloud infrastructure that can handle variable testing loads, and focusing on the specific use case of ten-year driving licence renewals that DVLA deals with.

CHAPTER 4 – EXPERIEMTS AND RESULTS

This chapter discusses the experiment setup, results analysis, and thorough evaluation of AWS Rekognition service effectiveness in comparing photos and checking photo standards for DVLA driving licence photo verification process. Checking the overall accuracy, fairness across different demographics, age progression analysis, and compliance with photo standards by using and filtering specified datasets for high quality images that fits defined criteria and evaluation metrics.

4.1 Experiment Setup

4.1.1 Comparison Request Processing Flow

Step 1: Starting a request

The process begins when a user sends a post request to the underneath endpoint with a JSON payload containing the request body.

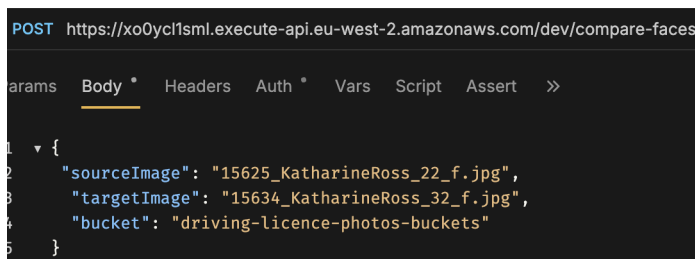


Figure 5 Verification Request Example

Step 2: API Gateway

It receives the request and performs some validations, converts it into a lambda event object and sends it to the compare faces lambda for processing.

Step 3: Lambda Execution

The lambda is triggered and starts processing it extracts the request input and checks if the required parameters like source image, target image and bucket are there if any are missing or don't exist in s3 a 400-status-code error is returned. Once we have these in place the lambda sends an API call to AWS Rekognition CompareFaces API with parameters source image, target image, SimilarityThreshold 95% and a QualityFitler set to high.

Step 4: AWS Rekognition CompareFaces API

AWS Rekognition CompareFaces API processes both images stored in S3 bucket and performs typical Facial Recognition process like performing facial features extraction and comparison algorithm, calculates the similarity percentage between the two faces extracts confidence score and returns a structured response to the lambda.

Step 5: Classification Processing

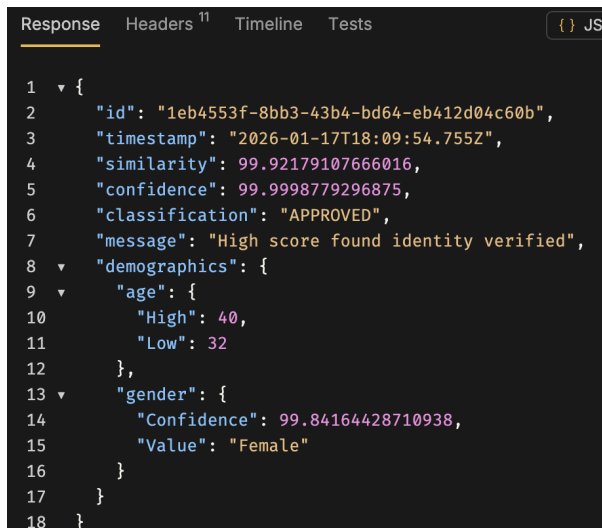
The lambda processes the Rekognition response thorough a simple three tire classification function if the Similarity score in the response is 95 or higher the photo is automatically approved as a true likeness. If the score is between 50 to 94 the photo is sent for manual review by a supervisor to double check it. If the similarity score is 0 or below 50 the photo is sent to fraud investigation team.

Step 6: Store Results in DynamoDB

The lambda stores the results in DynamoDB table including attributes like request id, timestamp, similarity, confidence, age, gender and classification.

Step 7: Reponses Generation:

The lambda returns a structured JSON response containing the following in blow screen.



```

1  {
2    "id": "1eb4553f-8bb3-43b4-bd64-eb412d04c60b",
3    "timestamp": "2026-01-17T18:09:54.755Z",
4    "similarity": 99.92179107666016,
5    "confidence": 99.9998779296875,
6    "classification": "APPROVED",
7    "message": "High score found identity verified",
8    "demographics": {
9      "age": {
10       "High": 40,
11       "Low": 32
12     },
13     "gender": {
14       "Confidence": 99.84164428710938,
15       "Value": "Female"
16     }
17   }
18 }

```

Figure 6 Verification Response Example

Step 8: Error Handling

The lambda has a comprehensive error handling covering invalid parameter request, image too large, image not found in s3 bucket, Api rate limit exceeded, invalid image format, throttling and server error.

Step 9: Monitoring and Logging

Throughout the entire process CloudWatch captures lambda logs and performance metrics, API Gateway request and response, Rekognition API calls and processing time and DynamoDB metrics.

Photo Standards Processing Flow

Step 1: Starting a Request

Similarly to the comparison process user sends a post request to the endpoint with the payload.

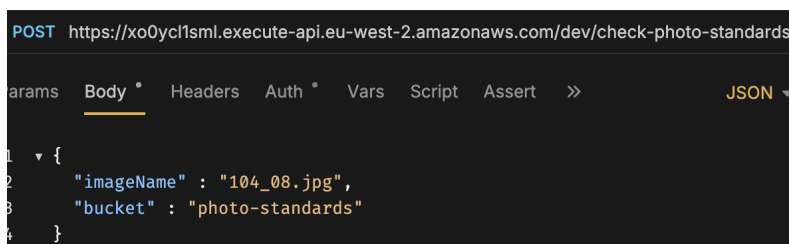


Figure 7 Photo Standards Request

Steps 2-3: API Gateway and Lambda Execution

This follows the similar pattern as above, routing the request to the check photo standards Lambda function which runs required validation.

Step 4: Rekognition API Calls

Following a similar process DetectFaces and DetectLables APIs are called to detect smiles, eyes, mouth open, wearing accessories etc.

Step 5: Verify Standards Function

This function has a comprehensive compliance checks it checks if there's multiple faces, facial expression, eye detection, pose analysis and quality assessments.

Step 6: Classification

Uses 2 classification if the photo meets all the requirements the photos is classified as Approved is there 1 or more violations like smiling wearing a hat etc the photo will be Unapproved.

Steps 7-9: Storage, Response and Monitoring

Similarly to the other lambda it stores results in a separate DynamoDB table, return response with detailed compliance checks, and follows the pattern for error handling and CloudWatch logging.

4.2 Results and Analysis

4.2.1 Analysis of Demographic Performance

This section aims to provides performance metrics covering the accuracy of CompareFaces API across different demographic groups. 359 genuine comparison pairs that are spread out across gender, ethnicity, and age groups are tested and analysed to test fairness and possible bias patterns in a somewhat a diverse population.

4.2.1.1 Overall Performance

The proposed biometric comparison application produced excellent results on all evaluations metrics as shown across all the tables in this section. The application achieved average similarity score of 99.92% with 99.98% confidence level as shown in below table. This high score exceeds industry standard for facial recognition accuracy.

Similarity	99.92%
Confidence	99.98%
TAR	100%
FAR	0%
FRR	0%

Table 3 Overall Performance

4.2.1.2 Security Analysis

To test the API robustness 30 pairs of imposters were used and the results are shown in below table. All pairs of impostors were correctly rejected, showing that the API is excellent at stopping fraud. Rejection of all imposter pairs proves the API ability at finding genuine and imposters, which provides security assurance for operational deployment.

Impostor Pairs Tested	30
TRR	30
FAR	0
Similarity	0%

Table 4 Security FAR

4.2.1.3 Gender Performance Analysis

Demographic analysis revealed a tiny performance variation between different genders, with female achieving slightly higher accuracy rates averaging 99.94%, while it was 99.90% for males. This means that females had a + 0.04% performance advantage, which falls within an acceptable statistical variance.

Gender	Total Pairs	Average Similarity	Confidence	Performance Variance
Female	179	99.94	99	0.08%
Male	180	99.90	99	0.12%
Overall	359	99.92	99	0.10%

Table 5 Gender Comparison

Although both male and female achieved an exceptional similarity score averaging above 99%. The tiny difference in performance between genders may suggest that the CompareFaces is more stable when processing female facial features. This could be because of the way the training data is distributed or because the algorithm is biased toward certain facial feature patterns. The fractional variance in male groups means that there is more inconsistency between different demographic subgroups which requires careful monitoring during operational deployment, so customers get the same level of service.

4.2.1.4 Ethnic Group Performance Analysis

CompareFaces worked well across ethnic groups, and the differences in performance stayed within the acceptable bias range less than 1% which is good for government biometric applications. Also, the maximum performance gap of between the highest and lowest scores e.g. Asian and Mixed-Race groups was less than 1% which is acceptable bias limits. Even so, some demographic subgroups had performance variations between different age groups but stayed within an acceptable range as well.

Ethnic Group	Sample Size	Average Similarity	Rank
Asian	10	99.99	1st
Middle Eastern	12	99.99	1st
Hispanic/Latino	33	99.95	2nd
White	168	99.93	3rd
Black	62	99.92	4th
East Asian	15	99.86	5th
Mixed Race	59	99.80	6th

Table 6 Ethnic Groups

Asian and Middle Eastern populations had the best performance averaging 99.99%, even though their sample size was the smallest. This suggests that the CompareFaces API is good at

recognising Asian and Middle Eastern facial features. The largest sample size, 168 pairs, comes from white populations, and their results are the most statistically sound, with 99.93% accuracy across a wide range of ages. Mixed Race populations have an overall performance of 99.80%, which is acceptable.

4.2.1.5 Female Demographic Analysis

Female populations across seven ethnic groups showed strong performance with interesting age-related trends. Testing 179 female pairs produced a 99.94% average similarity with just 0.02% variance between age groups, demonstrating reliable performance across different demographics.

Ethnic Group	16-30	30-40	40-50	50-60	60-70	70+	Overall
Black Females	99.95	99.63	99.99	99.86	99.87	99.91	99.87
White Females	99.995	99.89	99.99	99.93	99.93	99.91	99.94
Asian Females	99.98	99.99					99.98
South Asian Females	99.98	99.99	99.99				99.98
Mixed Race Females	99.77	99.99	99.98	99.99			99.93
Hispanic Females	99.99	99.99	99.75				99.91
Middle Eastern Females	99.97	99.99	99.96	99.999			99.98

Table 7 Female Demographics

Black females scored an average of 99.87% and had some variation across different age group. Performance declined in the 30-40 group averaging 9.63% and recovering at 40-50 years scoring 99.99%. This suggests facial features may be less stable during the thirties as facial features change during this period.

The largest group was white females, who had an average score of 99.94% across 91 pairs. This accuracy was consistent across age groups, with the highest rate of 99.99% between the ages of 40 and 50 and the data availability for this group ensured all ages are covered unlike others in this study.

Asian females had the highest scores 99.99% on average, even though there were only three pairs to compare. Their performance stayed consistently great in the 18–40 range, but we need more data from older age groups to be sure these results accurate across all ages.

Southeast Asian females with 7 pairs also did very well, with an average score of 99.98%. Their performance got better as they got older. The 40–50 group got almost all the answers right 99.99%, which indicates the API works well with older South Asian females.

Mixed-race females 31 pairs scored 99.77% for ages 18 to 30 but got a lot better 50 to 60 group achieving 99.99%. This indicates the API works well with mixed race females, but more data is needed to represent older groups.

Overall, Hispanic females did well 99.91%, but there was a drop in the 40 to 50 age group averaging 99.75%, so it needs to be investigated more closely despite the low data for this group.

Middle Eastern females with 9 pairs, achieved an average of 99.98%. The 30 to 40 and 50 to 60 groups both got almost perfect scores 99.99%, which indicates the API works very well with Middle Eastern facial features.

The study shows that most female ethnic groups have very high recognition rates, over 99.9%. The small overall difference shows that the API works the same way for all females, but some ethnicity age groups, may need more attention during deployment.

4.2.1.6 Male Demographic Analysis

Like Females the performance of male populations had a slight variation, with specific ethnic groups facing more age-related challenges. Although this is a fractional difference, but a more careful approach is needed to ensure services are delivered fairly.

The analysis shows performance patterns that are specific to certain demographics. These patterns help set priorities for algorithmic optimisation and operational deployment. White males, who make up the largest male group with 77 pairs and an average similarity of 99.92%, cover a wide range of ages, from 18 to 70 and older.

Ethnic Group	18-30	30-40	40-50	50-60	60-70	70+	Overall
Black Males	99.94	99.995	99.99	99.99	99.99	99.96	99.98
White Males	99.96	99.84	99.99	99.93	99.93	99.91	99.929
Asian Males	99.99	99.99	99.99				99.99
East Asian Males		99.99	99.99	99.23			99.74
Mixed Race Males	99.99	99.99	98.66	99.997			99.66
Hispanic Males	99.99	99.99	99.99	99.995			99.99
Middle Eastern Males		99.99	99.99	99.997			99.99

Table 8 Male Demographics

Black males stood out as the most reliable group 39 pairs, 99.98% average. Their scores barely fluctuated across age ranges, peaking at 99.998% for the 40-50 group.

White males with 77 pairs achieved good scores but declined in 30 to 40 group averaging 99.839% recovering in the next group 40 to 50, 99.997%. This decline pattern was in female data, indicating that facial features during this life stage change.

Asian averaged 99.99% and testing covered ages 18 to 50. The high score indicates the API handles their features well, but more data is required to cover other age groups. East Asian 50 to 60 group averaged 99.23%. With just 8 pairs total and a 99.74% overall average.

Mixed Race males delivered the worst performance in the 40 to 50 age group scored a dismal 98.656% across 28 pairs and averaged 99.66%, but that middle-age.

Hispanic and Middle Eastern males both excelled, scoring 99.99% and 99.99% respectively, though small sample sizes 3 pairs limit how much we can read into these results.

Although the average score was 99 for all groups the data exposes some issues East Asian in their fifties and Mixed Race in their forties both scored low compared to others could require motorring as these gaps could create discriminatory outcomes and need fixing before any widespread rollout.

4.2.1.7 Age Performance Analysis

Age	Female	Male	Combined
18-30 years	99.94	99.97	99.95
30-40 years	99.92	99.97	99.94
40-50 years	99.94	99.80	99.87
50-60 years	99.94	99.85	99.89
60-70 years	99.90	99.96	99.93
70+ years	99.91	99.93	92

Table 9 Age Analysis

Cross-demographic age analysis shows both universal patterns and challenges that are unique to certain groups. This data assists in establishing scope for deployment strategies. The comprehensive age analysis reveals trends spanning gender and ethnicity, while also emphasising a fractional demographic weakness that may not require focused interventions.

Looking at performance across different age groups reveals some interesting patterns and a few concerning trends that may require attention.

The youngest group 18-30 performs well at 99.95% combined, with Male scoring slightly higher than female. This flips in the thirties where both genders maintain good results around 99.94%, though female edge slightly.

At forties the combined average drops to 99.87% as males decline to 99.80% while females hold steady at 99.94%. This 0.14% gender gap is due to Mixed Race and South Asian males with lower scores at age period.

The 50 age group achieved 99.89% with males still lagging at 99.85% vs 99.94% for females, but by the sixties, the combined average reaches 99.93% as males recover to 99.96% outperforming females again. This recovery continues into the seventies where the combined score hits 99.92%.

What's particularly eye catching is how well elderly people perform. The 70+ achieved 99.92% combined, which contradicts the assumption that facial recognition accuracy declines with age. This indicates that the model training data included appropriate older faces, helping the API handle age-related facial changes effectively. Overall, middle-aged males from certain ethnic backgrounds face recognition challenges that may need addressing.

4.2.2 Age Gap Impact Analysis

The age gap analysis looks at how well the CompareFaces API works across different time gap between photo pairs. This is important to get an insight of how well it works for a ten-year renewal scenario where a time gap between photos can be 10 years. The assessment includes three separate age gap categories, showing patterns that are specific to each gender and traits that are important for long-term identity verification applications.

Gap	Male	Female	Combined	Difference
5-9 years	99.90	99.83	99.86	+0.07 Male
10 years	99.86	99.74	99.80	+0.12 Male
11+ years	99.94	99.61	99.78	+0.33 Male
Average	99.90	99.73	99.81	+0.17 Male

Table 10 Gap Analysis

The age gap test demonstrates CompareFaces resilience at handling various age gaps, with average similarity scores consistently achieving scores above 99.6% even for the longer periods. Male had higher scores across all age gap categories, with the performance difference increasing in extended time periods.

4.2.2.1 Age Gap 5–9 Years

This shorter 5 to 9 gaps are examined as lots of drivers change their photos before the ten-year renewal. The average similarity for males was 99.90%, while for females it was 99.83%. Males groups had a small advantage of + 0.07%, which is within the normal range of performance variation.

This time shows the best algorithmic performance for both males and females, which suggests that changes to the face that happen over the course of a decade don't make it much harder the CompareFaces API this likely due to the training dataset used. The tiny difference in similarity between genders indicates that the API works well for both.

4.2.2.2 10 years Age Gap

The ten- years age gap is the standard time between UK driving licence renewals where a change of photo is mandatory. It gives performance data for a common licence renewal operation. The average similarity for males remained at 99.86%, while the average for females dropped to 99.74%. This gave men a +0.12% advantage.

The drop in performance seen in both males and females is a normal part of aging over a full decade, but it is still very small and well within reasonable practical limits. The increased

gender similarity difference although its fractional but may indicates that female facial features may be more vulnerable to age-related recognition difficulties over extended periods.

4.2.2.3 Age Gap Analysis 11+

The 11+ years age gap category looks at how well the system works over extended periods, which can happen in rare cases of renewal or when historical verification is needed. The 99.94% average male score sets the highest performance. For females, it dropped to 99.61%, which was the lowest score and made the biggest gender difference of + 0.33%.

The unexpected rise in male similarity score for longer age gaps suggests ongoing aging may enhance facial distinctiveness for male traits, possibly due to more pronounced age-related alterations that augment algorithmic discrimination capacity. On the other hand, the significant drop in female performance 99.61% is close to the lower end of the optimal accuracy ranges, although it's within an acceptable range but might be challenging for females identity verification over a long period of time.



Figure 8 Comparison data sample

4.2.3 Photo Standards Compliance Analysis

The DetectFaces and DetectLabels APIs are used to check for photo standards compliance for photos that didn't meet defined set of standards. A total of 87 images in seven violation categories to see how effectively the APIs could find photos that fail to satisfy defined requirements.

4.2.3.1 Overall Performance of Photo Standards

The APIs worked differently for different types of violations. It was excellent for certain violations but had a lot of issues with others. The full evaluation gives important information about how reliable the APIs are for checking compliance with photos automatically in real-world DVLA settings.

Category	Number of Photos Tested	False Positives	FAR
Smile & Mouth Open	25	10	40%
Multiple Faces	4	0	0%
Glasses	9	0	0%
Face Occlusion	8	4	50%
Headwear	14	0	0%
Eyes	8	1	12%
Pose	12	0	0%

Table 11 Photo Standards

4.2.3.2 Perfect Detection Categories Analysis

The API was excellent at finding photos with multiple faces in the photo proven by its flawless performance on four test images, with a 0% false acceptance rate. This perfect record is essential for driving licence applications.

Glasses detection also worked perfectly on all 9 test images, correctly finding and labelling different violations. The 0% false accept rate shows that the API can reliably detect glasses that hide facial features needed for biometric comparison. This performance is important for DVLA and required for identity verification.

The API was good at detecting headwear and showed good results across 14 test images, with a false accept rate of 0%. It was able to correctly identify different types of head coverings, such as hats, kippah, turbans, headscarf, and hoods. Although the DetectLabels incorrectly labelled the hijab as a hoodie in some cases this makes the need for careful considerations as there are exemption for religious reasons. Overall, The APIs was able to detect various headwear types indicates comprehensive training data for some cultural and religious head coverings while adhering to security protocols. However, further training on the different types

of head covering such as a hijab may require more testing data to examine this further to ensure the API is fair for all cultures.

Pose detection worked perfectly on 12 test images, correctly identifying photos where the subjects are not facing forward or have their heads at the wrong angle. The 0% false accept rate shows that pose violations can be detected with confidence.

4.2.3.3 Problematic Detection Categories Analysis

The API worked moderately well with finding closed eyes, with a 12.5% false accept rate across 8 test images but failed to find closed eyes. The one false positive suggests that the API is somewhat reliable, but it may have trouble detecting certain eye states because of borderline cases or limitations in the training data which can impact the deployment. This level of performance needs human supervision for cases of borderline eye conditions.

The API was poor at detecting smiles and open mouth and scored 40% false acceptance rate across 25 test images. This was the largest violation category tested. The high false acceptance rate shows that it is hard to tell the difference between neutral expressions and smiling or open mouth. This variations in performance indicate that deploying this to critical government application can cause lots of issues and the manual verification process might be best to check for photos standards.

The API had the worst performance at detecting face occlusion with a 50% false accept rate across 8 test images. The high false acceptance rate indicates that the API struggles at finding partial face coverings that block facial visibility this could be down to poor training data. This level of performance is risky since it makes photos that fail to satisfy the requirements pass through the verification process.



Figure 9 Data Sample (Photo Standards)

4.2.4 Serverless Performance Analysis

CloudWatch monitoring revealed that the serverless application performed well during testing stage as shown in below table. The requests average processing time was 279.49 ms and the response times was between 1.6ms and 2,881.89 ms.

The slow processing time was mostly due to cold start as analysis revealed initialisation average was 272.80 ms. This represents majority of the total processing time. The implementation initialises AWS service clients inside the lambda function which is not ideal and possibly contributing to this and can be initialised outside the handler to allow reuse of connections for all invocations to reduce cold start issues.

Memory utilisation was consistent averaging 87.35 MB for all executions, staying within allocated memory size. The application processed the experimental workload without performance degradation, demonstrating the architecture's abilities to handle DVLA's operational requirements. However, due to time scale of this project it was not possible to load test the applications with appropriate load tests similar to the kind of traffic we get at DVLA online services.

Processing Average	Value
Processing Time	279.49 ms
Minimum	1.6 ms
Maximum	2,881.89 ms
Average Cold Start	272.80 ms
Cold Start Impact	97.6%
Average Memory Usage	87.35MB

Table 12 Serverless Performance

4.3 Discussion

This chapter examines the experimental results from results and analysis section, assessing the effectiveness of AWS Rekognition for DVLA Ten-Year-Renewal applications identity photo verification. The discussion examines the results and evaluates practical deployment issues found in photo verification and checking photo standards validation.

The CompareFaces API produced excellent results, which shows it may be ready to be used for identity verification. The accuracy levels were higher than what is expected in various applications. This may be due to AWS training datasets and advanced architectures.

The breakdown of the demographics shows some patterns that are worth looking into more closely. There is a small gender gap of +0.04% in favour of female subjects. This is not statistically significant, but it is better to what researchers have found in other facial recognition models and deployed systems. The cloud comparison research showed somewhat similar patterns on different commercial platforms, but Rekognition's gaps are much smaller than other providers. The difference is more telling: male performance changes more 0.12% than female performance 0.08%.

This suggests that the algorithm doesn't always work the same way for male facial features across different groups. This may be due to things like differences in facial features, different aging patterns, or even how the training data was spread out or perhaps due to the sample difference between the genders in this experiment.

The age progression findings are particularly important for DVLA's ten-year renewal scenario. Rekognition maintained strong similarity scores for images taken 5-9 years apart 99.86% and even at the full 10-year mark 99.80%, which demonstrates solid resilience against normal aging. However, female performance declined for gaps beyond 11 years dropping to 99.61%, hinting at limitations when the time gap extends further. This makes sense aging compounds over time through skin texture changes, volume loss, and structural shifts.

Studies have consistently demonstrated that female faces exhibit aging in subtle ways that algorithms find challenging to monitor, particularly when considering variations in makeup, hairstyle, and weight that differ more significantly than male patterns.

The lowest average in demographic was in East Asian men aged 50 to 60 dropped to 99.23%, and the percentage of Mixed Race male aged 40 to 50 dropped significantly to 98.65%. These results are in line with what we already know about problems with facial recognition: groups

that aren't well-represented in training data don't do as well. When mixing up race, gender, and age, creates complicated feature spaces that need a lot of examples to learn how to use them correctly. Mixed-race groups have struggled as the algorithm has to learn features that cross different ethnic groups.

Even with these differences, all demographic groups stayed above 98% accuracy, which is good enough for many uses. But government identity verification works differently. Security and fairness are essential. Since the sample sizes were different for each demographic group in this experiment, the results suggest that Rekognition should be deployed as an assistive tool for caseworkers to help them reach a decision rather than an independent decision-maker.

Caseworkers have full decision-making power for borderline cases or cases from demographic groups with a lot of variation. They can use it to help them make a decision. The trial phase would let the DVLA test it on a larger data and more diverse group of people extracting statistics on its performance and get feedback from caseworkers, to find issues that weren't expected before employing it. This hybrid approach values the technology's ability to consistently analyse features and the human ability to deal with edge cases, or unusual situations that algorithms have trouble with.

Photo Standards Validation Limitations

The results indicates that AWS Rekognition's DetectFaces and DetectLabels APIs have significant limitations for automated photo compliance validation. The clear distinction between excellent detection of some violations and poor detection of others demonstrates that there are essential differences in how Rekognition's training data and detection algorithms represent features.

Rekognition is excellent at finding different, high-contrast features with clear geometric or object-based characteristics because it has perfect detection rates for detecting multiple faces, different types of glasses, various headwear although it labelled hijab as a hoodie in various photo and pose violations. Excellent performance at detecting these features probably benefit from being well-represented in training datasets as they are common detection tasks in many applications other than identity verification.

However, the poor performance on smile detection 40% FAR, occlusion detection 50% FAR, and eyes detection 12.5% FAR exposes critical gaps for DVLA photo standards validations. To detect smiles and neutral expressions, you need to fully understand how facial muscles are arranged and how small changes in mouth position can affect them.

The 40% false acceptance rate suggests that Rekognition's smile detection may not be set up the way governments standards require it to be. Separating neutral expressions from slight smiles is something that human may struggle with.

Major issue like 50% false acceptance rate for occlusion detection, that impacts security and biometric matching. High error rate could be due to occlusion scenarios are complex, with things like hair or hands partially covering or shadows and lighting artefacts training may not have focused on subtle occlusion patterns, especially those that lack noticeable things like sunglasses or masks.

These limitations make it impossible for DVLA applications to fully automate the validation of photo standards. The UK government photo guidelines explicitly prohibit all of the above the Rekognition failed to detect. Since of the high false acceptance rates for these essential requirements, photos that fail to satisfy the requirements would often pass automated checks. This will lead to issues after deployment with biometric matching, exposing system to vulnerabilities and inconsistencies in standards.

The technical limitations show that there exists a basic mismatch between general-purpose computer vision APIs and specific compliance needs. The findings may indicate that the two APIs are made for general facial recognition where strict government standards don't have any significance. Rekognition's occlusion detection also looks for clear obstructions, like masks, not subtle partial coverings like hair or shadows that break photo rules. The API wasn't trained to follow government photo standards guidelines.

The findings suggest the DVLA should continue its current manual verification for standards compliance. Alternatively, DVLA could custom models that are trained on their photo database and optimised for ICAO standards. AWS Rekognition offer training a Custom Labels with specific need without having to build ML infrastructure from scratch, but this will require lots of training data and huge cost for running the model per hour which is against the current business requirements as currently cloud services are mostly used per use.

4.4 Chapter Summary

This covered the experimental setup and results for testing AWS Rekognition APIs performance for DVLA driving licence identity photo verification. The evaluation employed various metrics similarity scores, confidence levels, True Accept Rate, False Accept Rate, False Reject Rate, and demographic differentials across 359 genuine comparison pairs and 30 impostor pairs from diverse demographic groups.

The CompareFaces API worked well, with an overall similarity of 99.92% and 99.98% confidence. It also, correctly rejected imposter pairs without any false acceptances. The demographic breakdown did show some trends, though. There is a small gender gap of +0.04% in favour of females, and the performance of males varied more 0.12% than that of females 0.08% across different subgroups.

Testing age progression showed that the APIs was excellent with gaps of 5 to 10 years. However, after 11 years, female's performance dropped a little 99.61%. The real concerns came from specific demographic combinations East Asian males aged 50-60 dropped to 99.23%, and Mixed-Race males aged 40-50 fell to 98.66%. These results mean careful monitoring is essential if this gets deployed operationally. Overall, it's an excellent performance despite the fractional variations across different demographics.

Photo standards validation told a completely different story. DetectFaces and DetectLabels APIs did well on certain violations like multiple faces, glasses, headwear, and pose all got caught perfectly 0% FAR. But struggled with smile detection 40% FAR, occlusion detection 50% FAR, and eyes detection 12.5% FAR. These high false acceptance rates for critical DVLA photo requirements basically rule out using these APIs for automated photo standards checking.

The findings point toward using Rekognition as a support tool for caseworkers when comparing faces collecting data of comparisons then data can be analysed to decide on whether to automate it based on findings later down the line. Photo standards checking needs to stay manual for now, though DVLA might want to explore building custom models trained specifically on their task to achieve good results on all checks.

CHAPTER 5 – CONCLUSION, LIMITATION, AND RECOMMENDATIONS

5.1 Summary

This research tackled the challenge of automating DVLA photo verification for Ten-Year-Renewal driving licence applications. Manual photo verification can slow application processing and is prone to errors, comparing photos taken ten years apart is challenging as people get older. The study evaluated AWS Rekognition APIs effectiveness thorough a serverless application built with Lambda, S3, DynamoDB, and API Gateway.

Firstly, the literature was extensively analysed on existing facial recognition, age progression, demographic fairness, and ISO/ICAO standards. This review showed that we don't fully understand how commercial systems like AWS Rekognition deal with age gaps of ten years or more and work with different groups of people.

Secondly, set up methods, measurement metrics, extract and filter data from datasets like AgeDB, MORPH, CACD, FGNET, FRLL, CelebA-HQ for both comparisons and photo standards.

Thirdly, create a serverless application calling three Rekognition APIs like CompareFaces for biometric comparison, DetectFaces and DetectLabels APIs for photo standards validation.

Fourthly, test a total of 359 genuine pairs and 30 imposter pairs and a selection of compliant and non-compliant photos to test standards and analyse fairness and accuracy in both categories.

Finally, explore and analyse processing time and how well it would work in real life for DVLA operations.

5.2 Conclusion

AWS Rekognition's CompareFaces API achieved exceptional results for identity verification. The average similarity score was 99.92% and correctly rejecting all impostor pairs. Performance remained strong across age gaps, maintaining 99.80% similarity for important 10-year renewal period. The findings prove AWS Rekognition's API appropriateness for biometric matching.

However, demographic analysis revealed fractional but notable variations. While all groups achieved an average of over 98% accuracy levels, Mixed Race males 40 to 50 age group declined to 98.66% and East Asian males 50 to 60 average declined to 99.23% the lowest scores observed. On the other hand, female performance showed less variance 0.08% compared to males 0.12%, Indicating more consistent algorithmic behaviour across female demographics. The +0.04% gender differential favouring females, though small, warrants monitoring in operational deployment.

Photo standards validation exposed critical limitations. DetectFaces and DetectLabels APIs performed well for multiple faces 0% FAR, glasses 0% FAR, headwear 0% FAR, and pose violations 0% FAR. But failed on smile detection 40% FAR, occlusion detection 50% FAR, and eyes detection 12.5% FAR. High error rates for photo standards makes automating the process impossible for DVLA deployment without human oversight.

Based on these findings, Rekognition works best as a support tool rather than making decisions on its own. Caseworkers would see the similarity scores and confidence levels next to the photos. Caseworkers would still have the final decision to make, but the algorithm's consistent analysis would help them. This kind of hybrid approach makes sense and a lot safer until more analysis from review is gathered to reach a final decision. This hybrid approach balances caseworkers judgement, experience and Rekognition excellent patterns.

5.3 Limitation

Limitations include dataset was not big enough to get initial results, especially different demographic groups representation, some categories and populations had relatively small sample sizes which makes it harder to draw firm statistical conclusions about how well Rekognition works for these groups. Similarly, the age distribution wasn't balanced. There weren't many samples for people over 70, which means we can't reach a conclusion as to how the APIs handles elderly faces. Also, the 2021 census showed how diverse the UK population is and there are many ethnic minorities that were not represented in this study due to lack of inclusion in used datasets.

Another issue is that the testing involved a controlled benchmark datasets with controlled image quality and scenarios. But in the real world, DVLA receives photos from all sorts of professional photographers, automated photo booths, smartphones cameras etc. The lighting varies, backgrounds aren't always ideal, and quality can differ. It's quite possible that performance would drop when dealing with the with these situations.

There's also a timing consideration. This testing happened at one specific point using whatever version of Rekognition was current at the time. AWS keeps updating their algorithms, so what works today might perform differently six months from now. What we've got here is essentially a snapshot, not a long-term view of how the system behaves.

Looking ahead, future work should really focus on testing more diverse demographic data, especially for the groups where we saw some performance variation. Use a larger UK-specific dataset that's properly balanced would let researchers do more thorough fairness testing.

CHAPTER 6 – REFLECTIONS

This dissertation turned out to be more demanding than I'd expected, particularly balancing the practical work with academic research. My original timeline fell apart quickly the literature review took longer than planned because I kept finding relevant papers that needed to be included. Finding, preparation and filtering datasets took a while to ensure appropriate demographics balance.

Managing time properly became important. I had to figure out how to set goals that were achievable and sticking to deadlines. In some weeks I managed to lots done, other weeks much less. Breaking everything into smaller pieces helped, though I still got the timing wrong on some. This has been a great learning experience that helped improve my time management skills.

The technical work had a steep learning curve. I'd never worked with facial recognition technology before, so getting my head around algorithmic how it works, the different types of models and understanding bias issues, what makes training data good or bad, how different demographic groups get treated all of that was new to me. Then setting up the serverless was another challenge that helped me learn more about setting up AWS services as I went and gained valuable experience.

Working with benchmark datasets such as AgeDB and MORPH FGNET ACAD was unfamiliar. Understanding how they were built, what their limitations were, and how to filter them appropriately took considerable effort. The evaluation metrics FAR, FRR, demographic differentials were standard stuff in facial recognition research but entirely new to me, so I had to spend time figuring out what they meant and why they mattered.

My supervisor's input was valuable throughout this project. They pointed me toward useful academic resources on facial recognition and advice on methodology and how to structure the research properly which made a big difference in getting this done.

I have learned more about facial recognition, bias, datasets, and serverless. It's also shown me where I'm weak my stats could be better, and I need to understand the ethics side of government AI more. The hands-on skills from these designing experiments, making sense of results, getting AWS to do what you want should be useful down the line. Despite all the challenges, this has been a valuable experience.

REFERENCES

- [1] D. White and A. M. Burton, "Individual differences and the multidimensional nature of face perception," *Nature Reviews Psychology*, vol. 1, Mar. 2022, doi: <https://doi.org/10.1038/s44159-022-00041-3>.
- [2] P. P. Jonathon, G. Jeckeln, C. A. Hahn, A. N. Yates, P. C. Fontana, and A. J. O'Toole, "Unlocking the power of partnership: How humans and machines can work together to improve face recognition," *arXiv.org*, 2025. <https://arxiv.org/abs/2510.02570>.
- [3] J. P. Davis, K. Lander, R. Evans, and A. Jansari, "Investigating Predictors of Superior Face Recognition Ability in Police Super-recognisers," *Applied Cognitive Psychology*, vol. 30, no. 6, pp. 827–840, Aug. 2016, doi: <https://doi.org/10.1002/acp.3260>.^[1]_{SEP}
- [4] I. Adjabi, A. Ouahabi, A. Benzaoui, and A. Taleb-Ahmed, "Past, Present, and Future of Face Recognition: A Review," *Electronics*, vol. 9, no. 8, p. 1188, Jul. 2020, doi: <https://doi.org/10.3390/electronics9081188>.
- [5] "Serverless Architecture in Decisioning using AWS," *Iconic Research and Engineering Journals*, vol. 9, no. 4, Oct. 2025, doi: <https://doi.org/10.64388/irev9i4-1711066-1755>.
- [6] S. Anwarul and S. Dahiya, "A Comprehensive Review on Face Recognition Methods and Factors Affecting Facial Recognition Accuracy," *Lecture Notes in Electrical Engineering*, vol. 597, pp. 495–514, Nov. 2019, doi: https://doi.org/10.1007/978-3-030-29407-6_36.
- [7] R. V. Virgil Petrescu, "Face Recognition as a Biometric Application," *Journal of Mechatronics and Robotics*, vol. 3, no. 1, pp. 237–257, Jan. 2019, doi: <https://doi.org/10.3844/jmrsp.2019.237.257>.
- [8] S. Almabdy and L. Elrefaei, "Deep Convolutional Neural Network-Based Approaches for Face Recognition," *Applied Sciences*, vol. 9, no. 20, p. 4397, Oct. 2019, doi: <https://doi.org/10.3390/app9204397>.
- [9] M. Wang and W. Deng, "Deep Face Recognition: A Survey," *Neurocomputing*, vol. 429, Nov. 2020, doi: <https://doi.org/10.1016/j.neucom.2020.10.081>.
- [10] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," Apr. 2015. Available: <https://arxiv.org/pdf/1409.1556>
- [11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?" Accessed: Apr. 27, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/652cf38361a209088302ba2b8b7f51e0-Paper.pdf
- [12] H. Wu and K. W. Bowyer, "What Should Be Balanced in a 'Balanced' Face Recognition Dataset?," *arXiv.org*, 2023. <https://arxiv.org/abs/2304.09818>
- [13] A. Ribera Martínez, "The Debiasing Paradox: Facial Recognition Technology and Biometric Identification Systems in the Artificial Intelligence Act," *European Yearbook of Constitutional Law*, pp. 137–163, 2024, doi: https://doi.org/10.1007/978-94-6265-647-5_7.
- [14] T ashreef Abdullah Araf, A. Siddika, S. Karimi, and R. Alam, "Real-Time Face Emotion Recognition and Visualization using Grad-CAM," *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–5, Apr. 2022, doi: <https://doi.org/10.1109/ICAECT54875.2022.9807868>.

- [15] Ankita B, ChienChen H, L. P, and L. O, “Impact of Explainable AI on Reduction of Algorithm Bias in Facial Recognition Technologies,” pp. 85–89, May 2024, doi: <https://doi.org/10.1109/sieds61124.2024.10534745>.
- [16] P. Kittur, A. Pasha, S. Joshi, and V. Kulkarni, “A Review on Anti-Spoofing : Face Manipulation and Liveness Detection,” *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, pp. 1–6, Dec. 2023, doi: <https://doi.org/10.1109/cvmi59935.2023.10464680>.
- [17] De Mel, V. L. B., et al. "Survey of evaluation metrics in facial recognition systems. 2023" *ResearchGate* . Available at: https://www.researchgate.net/publication/372232460_Survey_of_Evaluation_Metrics_in_Facial_Recognition_Systems
- [18] M. Sundaram and A. Mani, “Face Recognition: Demystification of Multifarious Aspect in Evaluation Metrics,” *Face Recognition - Semisupervised Classification, Subspace Projection and Evaluation Methods*, Jul. 2016, doi: <https://doi.org/10.5772/62825>.
- [19] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments,” *inria.hal.science*, 2008. <https://inria.hal.science/inria-00321923>
- [20] P. Grother, M. Ngan, and K. Hanaoka, “Face recognition vendor test (FRVT) part 3: Demographic effects,” *Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*, vol. 8280, Dec. 2019, doi: <https://doi.org/10.6028/nist.ir.8280>.
- [21] K. Ricanek and T. Tesafaye, “MORPH: A Longitudinal Image Database of Normal Adult Age-Progression,” *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, 2006, doi: <https://doi.org/10.1109/fgr.2006.78> .
- [22] D. Deb, N. Nain, and A. K. Jain, “Longitudinal Study of Child Face Recognition,” *IEEE Xplore*, Feb. 01, 2018. <https://ieeexplore.ieee.org/document/8411226/>
- [23] V. Albiero, K. Bowyer, K. Vangara, and M. King, “Does Face Recognition Accuracy Get Better With Age? Deep Face Matchers Say No.” Accessed: Sep. 17, 2023. Available: https://openaccess.thecvf.com/content_WACV_2020/papers/Albiero_Does_Face_Recognition_Accuracy_Get_Better_With_Age_Deep_Face_WACV_2020_paper.pdf
- [24] Z. Huang, J. Zhang, and H. Shan, “When Age-Invariant Face Recognition Meets Face Age Synthesis: A Multi-Task Learning Framework.” Available: https://openaccess.thecvf.com/content/CVPR2021/papers/Huang_When_Age-Invariant_Face_Recognition_Meets_Face_Age_Synthesis_A_Multi-Task_CVPR_2021_paper.pdf
- [25] H. Yang, D. Huang, Y. Wang, and A. K. Jain, “Learning Continuous Face Age Progression: A Pyramid of GANs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019, doi: <https://doi.org/10.1109/tpami.2019.2930985>.
- [26] Z. Huang, S. Chen, J. Zhang, and H. Shan, “PFA-GAN: Progressive Face Aging With Generative Adversarial Network,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2031–2045, Jan. 2021, doi: <https://doi.org/10.1109/tifs.2020.3047753>.
- [27] S. Meena, K. Vats, and D. Kumar, “Review of factors affecting facial recognition algorithms performance,” *International journal of health sciences*, Jun. 2022, doi: <https://doi.org/10.53730/ijhs.v6ns3.8253>.

- [28] Eswar Sai Bokkasam, Blessy Manogna Danavulapadu, Dinesh Lakshmaiah, J. R. Enrathi, and Sudhakar Kothandapani, "Identification of identical twins using deep learning," *AIP conference proceedings*, vol. 3281, pp. 020019–020019, Jan. 2025, doi: <https://doi.org/10.1063/5.0247025>.
- [29] M. Scheuerman, J. Paul, and J. Brubaker, "How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services," vol. 144, 2019, doi: <https://doi.org/10.1145/3359246>.
- [30] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb. 2020, doi: <https://doi.org/10.1145/3375627.3375820>.
- [31] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and Reducing the Damage of Dataset Bias to Face Recognition with Synthetic Data."
- [32] D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–44, Apr. 2023, doi: <https://doi.org/10.1145/3494672>.
- [33] T. Gebru *et al.*, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: <https://doi.org/10.1145/3458723>.
- [34] G. Finocchiaro, "The regulation of artificial intelligence," *The regulation of artificial intelligence*, vol. 39, Apr. 2023, doi: <https://doi.org/10.1007/s00146-023-01650-z>.
- [35] V. L. Raposo, "(Do not) remember my face: uses of facial recognition technology in light of the general data protection regulation," *Information & Communications Technology Law*, pp. 1–19, Mar. 2022, doi: <https://doi.org/10.1080/13600834.2022.2054076>.
- [36] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019, doi: <https://doi.org/10.1038/s42256-019-0048-x>.
- [37] S. Wilkinson, "Artificial intelligence, facial recognition technology and data privacy," *Journal of Data Protection & Privacy*, vol. 3, no. 2, pp. 186–198, Jan. 2020
- [38] A. Cavoukian, "Privacy by Design The 7 Foundational Principles Implementation and Mapping of Fair Information Practices," 2009. Available: https://student.cs.uwaterloo.ca/~cs492/papers/7foundationalprinciples_longer.pdf
- [39] R. Xu, N. Baracaldo, and J. Joshi, "Privacy-Preserving Machine Learning: Methods, Challenges and Directions," *arXiv:2108.04417 [cs]*, Sep. 2021, Available: <https://arxiv.org/abs/2108.04417>
- [40] M. Ferrara, A. Franco, D. Maio, and D. Maltoni, "Face Image Conformance to ISO/ICAO Standards in Machine Readable Travel Documents," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 4, pp. 1204–1213, Aug. 2012, doi: <https://doi.org/10.1109/tifs.2012.2198643>.
- [41] "Photograph quality The photographs must be." Available: https://www.icao.int/Security/mrtd/Downloads/Technical%20Reports/Annex_A-Photograph_Guidelines.pdf
- [42] 14:00-17:00, "ISO/IEC 19794-5:2011," *ISO*. <https://www.iso.org/standard/50867.html>

- [43] “Photograph quality The photographs must be.” Available: https://www2023.icao.int/Security/mrtd/Downloads/Technical%20Reports/Annex_A-Photograph_Guidelines.pdf
- [44] H. P. Office, “Photo standards,” *GOV.UK*, Sep. 29, 2022. <https://www.gov.uk/government/publications/photographic-standards/photo-standards-accessible>.
- [45] J. Urbanus, Bille, and Y. Musa, “Software Framework for Face Recognition System Using Amazon Rekognition,” *International Journal of Computer Science and Mathematical Theory (IJCSMT) International Journal of Computer Science and Mathematical Theory*, doi: <https://doi.org/10.56201/ijcsmt.v10.no5.2024.pg135.154>.
- [46] M. Amir, Ioana Ghergulescu, and A.-N. Moldovan, “Review and Comparison of Facial Emotion Recognition Services and APIs,” pp. 1–7, Jun. 2025, doi: <https://doi.org/10.1109/aicconf64766.2025.11064077>.
- [47] S. Jha, M. Arora, Y. Sharma, A. Anand, and D. Sharma, “Comparative Analysis of Cloud Computing Based Face Recognition Services,” *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, May 2022, doi: <https://doi.org/10.1109/com-it-con54601.2022.9850927>.
- [48] Miryala, N. K. “Cloud Performance A Comparative Study of AWS vs. Azure,” 2024 *International Journal of Computer Engineering and Technology (IJCET)*, 15(2), 208-223.
- [49] K. Yang *et al.*, “Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets,” *The Visual Computer*, Jun. 2020, doi: <https://doi.org/10.1007/s00371-020-01881-x>.
- [50] S. Tariq, S. Jeon, and S. S. Woo, “Evaluating Trustworthiness and Racial Bias in Face Recognition APIs Using Deepfakes,” *Computer*, vol. 56, no. 5, pp. 51–61, Apr. 2023, doi: <https://doi.org/10.1109/mc.2023.3234978>.
- [51] A. Elivanova, “Privacy in Responsible AI: Approaches to Facial Recognition from Cloud Providers,” *arXiv.org*, 2025. <https://arxiv.org/abs/2503.04866> (accessed Aug. 01, 2025).
- [52] M. Ahmed, “UK passport photo checker shows bias against dark-skinned women,” *BBC News*, Oct. 07, 2020. Available: <https://www.bbc.co.uk/news/technology-54349538>
- [53] J. Rees, “Police force facial recognition use ruled unlawful,” *BBC News*, Aug. 11, 2020. Accessed: Aug. 12, 2020. [Online]. Available: <https://www.bbc.co.uk/news/uk-wales-53734716>
- [54] “Home Office drops ‘racist’ algorithm from visa decisions,” *BBC News*, Aug. 04, 2020. Available: <https://www.bbc.co.uk/news/technology-53650758>
- [55] L. R. Montgomery, “Facebook and the Biometric Information Privacy Act Litigation,” *Endnotes: The Journal of the New Members Round Table*, vol. 13, no. 1, p. 72, 2025, Accessed: Nov. 07, 2025. [Online]. Available: <https://journals.ala.org/index.php/endnotes/article/view/8492/11815>
- [56] Top Class Actions, “Google Photos face recognition privacy \$100M class action settlement,” *Top Class Actions*, Jun. 03, 2022. <https://topclassactions.com/lawsuit->

settlements/closed-settlements/google-photos-face-recognition-privacy-100m-class-action-settlement/

[57] Big Brother Watch, “Big Brother Watch Briefing on facial recognition surveillance,” 2020. Available: <https://bigbrotherwatch.org.uk/wp-content/uploads/2020/06/Big-Brother-Watch-briefing-on-Facial-recognition-surveillance-June-2020.pdf>

[58] Office for National Statistics, “Population of England and Wales,” *GOV.UK*, Dec. 22, 2022. <https://www.ethnicity-facts-figures.service.gov.uk/uk-population-by-ethnicity/national-and-regional-populations/population-of-england-and-wales/latest/>

[59] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” 2016. Available: https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVP_R_2016_paper.pdf

[60] “Photo style.” Available: https://assets.publishing.service.gov.uk/media/5a7e2f0140f0b62302689b37/Photoguidance_v7.pdf

[61] “Get a passport photo,” *GOV.UK*. <https://www.gov.uk/photos-for-passports>

[62] N. GANDHI, “AgeDB_Database,” *Kaggle.com*, 2019. <https://www.kaggle.com/datasets/nitingandhi/agedb-database>

[63] AAI, “FGNET Dataset,” *Kaggle.com*, 2024. <https://www.kaggle.com/datasets/aiolapo/fgnet-dataset>

[64] Paweł Dombrzalski, “CACD filtered dataset,” *Kaggle.com*, 2025. <https://www.kaggle.com/datasets/pdombrza/cacd-filtered-dataset>

[65] Chirag Sai Panuganti, “MORPH-2,” *Kaggle.com*, 2023. <https://www.kaggle.com/datasets/chiragsaipanuganti/morph>

[66] “CelebA-HQ resized (256x256),” *www.kaggle.com*. <https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256>

[67] L. DeBruine and B. Jones, “Face Research Lab London Set,” *figshare*, 2017, Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666

Project Gantt Chart

1		September				October				November				December			
2		Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
3	Project planning																
4	Literature review and tech evaluation																
5	Application design																
6	Implementation																
7	Testing and validation																
8	Refinement																
9	Report																
10	Documentation and submission																
11	Presentation																

Figure 10 Gantt Chart Proposal

	September				October				November				December				January			
	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4	Week 1	Week 2	Week 3	Week 4
Project planning																				
Literature review and tech evaluation																				
Application design																				
Implementation																				
Testing and validation																				
Refinement																				
Report																				
Documentation and submission																				

Figure 11 Gantt Chart